Toward Monetizing Data for AI-driven Services on Cloud Computing and Blockchain

Ahmed Saleh Bataineh

A Thesis

in

The Department

of

Concordia Institute for Information Systems Engineering

Presented in Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy (Information and Systems Engineering) at

Concordia University

Montréal, Québec, Canada

November 2021

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: **Ahmed Saleh Bataineh**

Entitled: **Toward Monetizing Data for AI-driven Services on Cloud Computing and Blockchain**

and submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy (Information and Systems Engineering)

complies with the regulations of this University and meets the accepted standards with respect to originality and quality. Signed by the Final Examining Committee:

_____ Chair
Dr. Chun-Yi Su

_____ External Examiner
Dr. Muhammad Younas

_____ Examiner
Dr. Juergen Rilling

_____ Examiner
Dr. Rachida Dssouli

_____ Examiner
Dr. Roch Glitho

_____ Supervisor
Dr. Jamal Bentahar

_____ Co-supervisor
Dr. Rabeb Mizouni

Approved by _____
Abdessamad Ben Hamza, Chair
Department of Concordia Institute for Information Systems Engineering

17/Nov/2021

_____
Mourad Debbabi, Dean
Faculty of Engineering and Computer Science

Abstract

Toward Monetizing Data for AI-driven Services on Cloud Computing and Blockchain

Ahmed Saleh Bataineh, Ph.D.

Concordia University, 2021

AI-driven services and data collecting-based applications are these days the talk of the town in the field of computer science and beyond. This is often obvious since one can effortlessly take note, by looking around, how this field has ended up fundamental in our day to day lives beginning from business intelligence and market analysis down to virtual personal assistants (e.g., Siri, Google Now) and social media analysis (e.g., people you may know on Facebook). However, the research communities anticipate a turn down in the revolution of AI-driven services and data collecting-based applications due to the deficiency within the accessibility of huge data that ought to be collected or (pre-)trained using machine learning technologies. Mainly, collecting and integrating the big complementary data scattered across foundations and countries entails high costs and management challenges associated with finding and getting on board multiple data providers. In this thesis, we tackle this problem by designing a data market platform on top of the cloud computing technology. The data market platform helps the data providers and data consumers find and meet each other, while the cloud technology provides the required computing resources to execute computational tasks associated with data processing. This thesis starts by searching and investigating the most efficient business theories to model the data market platform. As a result of our search, the two-sided market theory has been proposed as a successful underlying model to design the data market platform.

The two-sided market theory has been improved in this thesis to handle challenges associated with considering data as an economic good on the one hand, and reshaping the business of the cloud

computing to act as a data market platform on the other hand. Mainly, we introduce a novel game theoretical model (Two-sided game), which consists of a mix of cooperative and competitive strategies. The players of the game are the *big data providers*, *cloud computing platform*, and *data consumers*. The strategies of the players are modeled using the two-sided market theory that takes into consideration the network effects (externalities) among involved parties. The externalities refer to the mutual impact of the number of data providers and data consumers on each other. The objective of this game is to enable the cloud to be an active platform that can help big data service providers reach a wider set of customers and cloud users (i.e., data consumers) to be exposed to a larger and richer variety of data to run their data analytic tasks. The proposed game has been improved further to deliver complementary data services among multiple data providers over a cloud intermediary platform. More specifically, we formalize the problem as an extended two-sided market model by courting on one side some influential data providers in order to attract other data providers on the same side to form a bundling of data services. The final game aims to dynamically distribute the cloud computing resources among computational tasks of data providers to maximize the social welfare of all involved parties. The game has been supported by a mechanism to handle potential undesired behaviours such as the greedy and irrationality behaviours of involved parties. The game also provides a clear pricing mechanism that estimates the monetary value of data considering the actual need of the data consumers.

The thesis ends up by involving the blockchain technology in the process of monetizing data. The blockchain technology has recently proved to be an efficient solution for guaranteeing the security of data transactions in data trading scenarios. The benefits of the blockchain in this domain have been shown to span over several crucial security and privacy aspects such as verifying the identities of data providers, detecting and preventing malicious data consumers, and regulating the trust relationships between the data trading parties. However, the cost and economic aspects of using this solution such as the pricing of the mining process have not been addressed yet. In fact, using the blockchain entails high operational costs and puts both the data providers and miners in a continuous dilemma between delivering high-quality security services and adding supplementary costs. In addition, the mining leader requires an efficient mechanism to select the tasks from the mining pool and determine the needed computational resources for each particular task in order to maximize its

payoff. Motivated by these two points, we propose in this thesis a novel game theoretical model based on the two-sided market approach that helps both the data providers and miners determine the monetary reward and computational resources, respectively.

Acknowledgments

It is a genuine pleasure to express my deep sense of thanks and gratitude to my supervisors, Dr. Jamal Bentahar and Dr.Rabeb Mizouni. I am extremely thankful to my parents, my wife, brothers and sisters as well as my research lab colleagues for providing me necessary support. Furthermore, I would like to thank Concordia University for all the research facilities that have been provided tome to carry out this work.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In this chapter, we discuss the research context and problem statement, and formulate the research questions, consequently. We further summarize the PhD research objectives to be accomplished. The chapter ends by providing the thesis organization.

1.1 Motivations and Context of Research

Nowadays, data collecting-based applications and Artificial Intelligence (AI)-driven services are being used in many industries and sectors such as driver-less cars, medical care, finance, etc. Market analytics, as a data collecting-based application example, relies on collecting and combining personal data from multiple providers to study the consumer behavior and interest to determine the future directions and pricing processes for certain products. Retailers, banks and insurance firms are examples of stakeholders that collect and combine personal data to target their consumers with marketing offers and promotional products. AI-driven services, as another example, release technology solutions to assist organizations and individuals by executing machine learning and data analytics procedures on massive data involving multiple data types, generated by multiple data providers. For example, Riskified, an AI-driven recommendation service, helps e-commerce sites release new products and enter new markets as well as identify legitimate shoppers. Riskified required more than one billion past transactions including data about the products, stores, user's purchases, brands, and associated data about the customers to make excellent instant decisions. Such inherently combinatorial datasets, which are formed by integrating different data types from multiple data providers are

refereed to in the thesis as complementary data. In addition, data collecting-based applications and AI-driven services are refereed to as data consumers.

However, the research communities expect a turn down in the revolution of data collecting-based applications and AI-driven services (data consumers) due to the shortage in the availability of big complementary data that need to be collected or (pre-)trained using machine learning algorithms [106]. Specifically, data collecting-based applications and AI-driven services entail high costs associated with collecting and integrating the big complementary data scattered across foundations and countries. Moreover, finding and getting on board multiple data providers raise management challenges associated with the level of collaboration, participation and consensus among the data providers to deliver a bundle of complementary data. Similarly, the data providers entail high costs for marketing, publishing, and delivering their data services for a wide range of data consumers. This problem, i.e., finding big complementary data, is only further exacerbated once data collecting-based applications and AI-driven services concern the real-time factor. For example, ride-roads-map applications solve their prediction problems according to real-time traffic data collected from vehicles supported by IoT sensors. Therefore, it is imperative that there are real-time market structures for buying and selling data.

This thesis aims to design and develop a data market platform that efficiently gets on board the data providers (sellers) and data consumers (buyers) in order to exchange and share data commodities transactions. The data market platform is not the actual data owner, but it mainly aims to facilitate data monetization by introducing the providers and consumers to each other and performing computing tasks associated with data processing. Consequently, the data market platform has to own intrinsic monetary properties in terms of a wide social network of data consumers and providers, and powerful computing infrastructure. In this thesis, we argue that the cloud computing is the most appropriate technology that can serve as platform for monetizing data as it hosts an explosive amount of data coming from a variety of enterprises and manufacturers that are deployed on its computing platforms. For example, the study reported in [2] revealed that one million customers deploy their own enterprises on Amazon, spending 30 billion USD on persistent storage on Amazon EC2 instances and generating 600 ZB of data per year [46]. Thus, the cloud computing could be used to liberate AI-driven services from having to search and discover appropriate data providers

and to give them the opportunity to extract valuable patterns of information from massive complementary data, originating from multiple data providers. Such a proposal, i.e., the cloud computing as a data market platform, entails challenges associated with considering the data as a financial asset or an economic good from one hand, and reshaping the business of the cloud computing from the other hand. From the first hand, i.e., data as a financial asset, data sets can be replicated and shared among multiple data consumers at zero marginal cost. In addition, the data types in complementary datasets exhibit a range of correlations and dependencies in the sense that the availability of a certain data type impacts the monetary value of other data types. Moreover, the varieties of interests, scopes, and businesses of data-based applications and AI-driven services (data consumers) lead to non-uniform data monetary values. From the second hand, i.e., reshaping the business of the cloud computing, this computing paradigm will become a two-sided market intermediating the interactions between the data providers and data consumers. Consequently, the strategies of cloud computing resource allocation will be upgraded in such a way that maximizes the social welfare and revenues of involved parties (i.e., cloud computing, data providers, and data consumers). This thesis takes a holistic view of this problem and explores solutions combining concepts from economics of data, market design, applied game theory, resource allocation, and optimization under uncertainty.

Besides the significant need for such a data market platform to sustain and accommodate the revolution of AI-driven services and data collecting-based applications, such a market opens the door for multi-billions businesses involving buying and selling data. In parallel, research and industry communities are paying recently attention toward blockchain technologies that opt viable solutions to generate abundant, secure and complete raw data. As a result of this attention, the data contained in the blockchain ledger is expected to worth up to 20% of the global big data market and generate up to 100 billion dollars in annual income [1].

1.2 Related Work and Problem Statements

In this section, we will answer the following question: What is the research gap in the literature?, which will help us state our problem statement. The problem of collecting and monetizing data has been studied either in presence of the peer-to-peer model or the merchant model. Under

the peer-to-peer model, proposals such as [18, 44, 52] focus and discuss the relationship between organizations (i.e., second owners) and consumers without involving the actual data owners. Specifically, organizations (second data owners) benefit from the data of individuals by selling them while the actual owners are not appropriately compensated. Facebook, as a practical example, earned a total revenue of 3.85 billion in the fourth quarter of 2014 from ads [3] - a fact that allows marketers to reach the personal data of their users. Credit bureaux, such as Equifax, Experian and TransUnion, sell their consumers' personal data to retailers, banks, insurance firms, and government agencies. The federal agency for Medicare and Medicaid Services, another class of organizations that are engaged in the business of buying and selling personal data, sells medical claims that include medical, demographic, and geographic personnel data to third parties [4]. This model raises challenging issues related to privacy, and recently selling data to a third party is being seriously discussed by law communities. For example, the General Data Protection Regulation (GDPR)[1] calls for mandating the process of exchanging personal data by law to protect the fundamental right of actual data owners and allow them to have a control on their personal data. Such a call requires a solid and more coherent monetizing data platform backed by a strong market model to incorporate the primary data owners in the process of buying and selling data. Specifically, the monetizing data platform has to develop an effective mechanism to guarantee an adequate participation of these owners in the data monetizing process.

Under the merchant model [11], monopolistic organizations take the possession of seller's data, reprocess them, and resell them to consumers at retrial price. The merchant model for monetizing big amounts of data raises serious challenges related to revenue maximization and leads to an ineffective third party data monetizing schema. Specifically, studies adopting the merchant model such as [46, 48, 70] overlook its limitations that include high processing/operation costs once a merchant wants to achieve higher product verity in terms of type, quality and quantity of products. Moreover, under the merchant model, the data providers (actual data owners) aim to maximize their revenue from their data commodities while the third party (information service provider as third party) aims to minimize the cost of the raw data. In parallel, at the information consumer-side, the third party aims to maximize its revenue from selling the processed information while the consumer

[1] https://gdpr-info.eu/

aims to minimize the cost of information commodities considering the maximum available quality and quantity of information. The resulting equilibrium from such aggressive competitions among the involved parties leads to less and coarse distribution for total surplus. In addition, data differ from other economic goods for the possibility to be resold to many consumers at the same time. In the merchant model, because of selling economic goods that cannot be resold, the equilibrium of the market is given by the intersection of the demand and supply curves, which means the quantities of goods needed by the consumers equal the quantities of goods provided by the sellers. This does not hold in our case since the same data can be shared between all the consumers. For instance, if there are 1000 consumers and each consumer requires 100 units of particular data, it does not mean that 100000 data units are needed, where it is enough for the merchant to buy 100 units and share them with the whole data consumers. This will largely raise the competition between data providers and push them to accept lower prices, which negatively affects their total surpluses and leads to inadequate involving of providers in the process of monetizing data. This leads us to the first subproblem statement **P1**:

P1: *The state of the art associated with the data domain research has not studied yet the idea of data market where the data are shown as commercial commodities, and involved parties (i.e., data providers and data consumers) find, meet, and match each other effectively. Specifically, the merchant and peer-to-peer models that dominated the research proposals in this context have been basically proposed for formalizing the businesses of information service providers without addressing the complex issues associated with designing an open data market platform such as involving primary actual data owners and the monetary data pricing.*

The first subproblem raises the following research question: *what is the most optimal business model that can be used to design the data market platform?* To tackle this question, this thesis investigates the two-sided market theory as a successful business model for designing the data market platform. The two-sided market theory concerns getting on board two distinct groups of users, i.e., buyers and sellers, by a subsidizing mechanism. The subsidising mechanism attracts one of the groups, for instance the buyers, by zero transaction fees, which consequently attracts the sellers and incentivizes them to pay higher versus accessing the buyers. The subsidising mechanism efficiency mainly depends on the mutual impact of the number of buyers and sellers on each others. The

mutual impact of the number of buyers and sellers is refereed to in the thesis by the cross-group externalities.

The data market platform has to originally be trusted, well-known and socially rich in terms of number of data providers and consumers using it. In addition, the platform has to be equipped by a powerful IT infrastructure such as servers and virtual machines to run and deploy computational tasks associated with storing and processing the data. Consequently, the immediate question is: *what is the most suitable technology that can act as data market platform?* and obviously the intuitive answer for this question is *cloud computing*. In fact, the cloud hosts an explosive amount of data coming from a variety of enterprises and manufacturers that are deployed on its computing platforms. For example, the study reported in [2] revealed that one million customers deploy their own enterprises on Amazon, spending 30 billion USD on persistent storage on Amazon EC2 instances and generating 600 ZB of data per year [46]. This explosive amount of data generated and stored on cloud resources forms the backbone for Artificial Intelligence (AI)-driven services and opens the door for a new cloud business paradigm, enabling the latter to be an active platform for monetizing data that benefit these services. Motivated by this vision of the cloud as a data market platform, and the promising results drawn by investigating the two-sided market business model for data market platform, this thesis proposes to reshape the business model of the cloud computing in the presence of the two-sided market theory. The two-sided cloud computing platform for monetizing data for AI-driven services plays two key roles: 1) introducing wide social networks of AI-driven services to the data providers and vice versa; and 2) providing computing infrastructure for both the AI-driven services to deploy their machine learning and data analytics procedures and data providers to deploy their collected data.

The proposal of the two-sided cloud computing platform for monetizing data for AI-driven services raises challenges associated with distributing elastically the cloud computing resources and revenue maximization including the interactions among three entities: cloud computing platform, data providers, and data consumers. Few proposals in the literature touch the problem of cloud resources utilization and revenue maximization including the interactions among the three entities using both competitive and collaborative game-based models [60, 59, 57, 95, 23, 101, 33, 101]. In the proposals that adopt competitive games, cloud computing competes aggressively with the

involved parties to maximize its revenues from renting or selling computing resources. Such an aggressive competition leads to coarse distribution/provision for cloud resources and excludes the small service providers from the market. On the other hand, the proposals adopting collaborative models are not capable to provide an efficient mechanism to split the earned revenues. However, adopting traditional game theory concepts (e.g., Shapley value and Nash equilibrium) to distribute the revenue that results from the cooperation among the different parties suffers from several limitations when applied in dynamic data trading scenarios over the cloud. Specifically, 1) although such concepts might be highly efficient in scenarios wherein all the involved parties are rational, their effectiveness starts to decrease in the presence of parties that are heterogeneous and prefer to deviate from the equilibrium points. For example, recent studies have revealed that only 37% of the players tend to accept the Nash equilibrium in cooperative games (interested readers can consult behavioral games and ultimatum games [39] for further details); and 2) even though the Shapley value approach fairly splits the revenues among the cooperative entities based on their contributions, it becomes inapplicable in cases wherein the contributions of entities cannot be measured (which applies to the cloud scenario considered in this work). Specifically, the cloud provider adds an ethereal/intangible, yet significant, contribution to the coalition via introducing wide social networks of data consumers to those of data providers. Moreover, data providers own the data which forms the core of this new business. This creates a continuous dilemma between data providers and cloud providers about who makes the most significant contribution to the coalition and hence who deserves the biggest share of the revenues. Equal distribution, so-called *fifty-fifty*, is one approach to split the revenues between the cloud provider and data provider. However, as mentioned earlier, the rationality and greediness of the involved parties (i.e., the cloud provider and data providers) prohibit the success of such a strategy. This leads us to the conclusion that we are dealing with a behavioral and ultimatum game in which two players (proposer and responder) argue to split a certain amount of revenue. The proposer is endowed with a sum of revenue and is responsible for splitting this sum with the responder. The responder may accept or reject the sum. In the case the responder accepts the sum, the revenue is split as per the proposal; otherwise, both players receive nothing. This leads us to the second subproblem that will be handled in this thesis:

P2: *The cloud computing can play a critical role as a data market platform intermediating and*

facilitating the interactions between the data providers and data consumers. However, the state of the art associated with the cloud technology has not adequately addressed the cloud as an intermediate platform. Mainly, the underlying models and approaches concerning distributing elastically the cloud resources once the cloud acts as an intermediate platform are inefficient in terms of revenue maximization and distribution. This raises the following research questions: 1) How can the business of the cloud computing be reshaped to act as a data market platform?; and 2) How can the elasticity of the cloud computing resources be integrated with two-sided market theory?

As mentioned earlier, AI-driven services and data collecting-based applications (data consumers) require big chunks of different data types coming from multiple data providers, which form complementary datasets. The data types in complementary datasets exhibit a range of correlations and dependencies in the sense that the availability of a certain data type impacts the monetary value of other data types. These correlations and dependencies among different data types place the two-sided cloud computing and the two-sided market theory itself in front of a new challenge associated with pricing a certain data type in the presence of other data types participating in the same complementary dataset. This motivates the need for a bundling data strategy. However, the two-sided market theory is not equipped with a bundling strategy wherein the internal dependencies among users belonging to the same group have not been considered. On the other hand, bundling commercial goods and services are in general modeled by coalitions among the suppliers of these commercial goods. However, such coalitions are unpractical in the scenario of bundling data for two reasons: 1) the number of data providers participating in the complementary datasets is relatively large, where the actual data owners (i.e., individuals carrying IoT devices) are included in the data monetizing process; and 2) data providers exhibit diverge level of professionalism, irrationality, preferences and conflicts of interest. This bundling data challenge adds an orchestration role for the two-sided cloud computing to controls the supply of data services into the bundled data services. This leads us to the third subproblem:

P3: *How can we design a strategic game that aims to deliver complementary data services among multiple data providers over the two-sided data market platform?*

The Blockchain technology has been widely used as an intermediary paradigm to address the privacy and security challenges that arise during the data trading transactions. These challenges

include user authentication, data integrity guarantee, and providers' privacy preservation in various domains such as Internet of Things (IoT), data analytics, and mobile crowd-sensing. In the context of data trading using blockchain, three players are to be considered: miners, data providers and data consumers. Miners are responsible for supervising and regulating the execution of what is known as *smart contracts*. A Smart contract is a self-executing computer program that states and organizes the agreed terms of a certain data transaction such as the desired quality of service clauses and secure payment mechanism between the data providers and data consumers. The unprecedented wave of IoT demands on the blockchain technology makes this latter a rich platform of valuable data coming from a variety of providers. However, using the blockchain entails high operational costs and puts both the data providers and miners in a continuous dilemma between delivering high-quality security services and adding supplementary costs. Furthermore, the mining leader requires an efficient mechanism to select the tasks from the mining pool and determine the needed computational resources for each particular task in order to maximize its payoff. In the literature, there is lack of attention on the aspect of the business model that would enable data trading over blockchain where the main stream research in the general context of data focuses on developing mechanisms of data resource management. Several challenging issues are yet to be addressed because of the lack of attention on the business model that would enable data trading over blockchain. The key challenges are assigning optimal amount of computational units to the mining tasks, sustaining optimal payoffs to involved players and serving data requests on time. This leads us to the fourth subproblem of this thesis:

P4: *How can we design a strategic game that aims to assign optimal amount of computational units to the mining tasks, sustaining optimal payoffs to involved players and serving data requests on time?*

1.3 Contributions

(1) **Contribution 1**: to resolve the subproblem **P1**, we design a data market platform using the two-sided market theory. The proposed two-sided data market platform enables providers (i.e., primary data owners) and data consumers to meet each other and perform data trading

transactions. The proposed underlying theory of data market platform concerns a subsidising mechanism to get on board the data providers and data consumers. The subsidising mechanism attracts one of the groups, lets say the data consumers, by low transaction fees, which consequently attracts the data providers and incentivizes them to pay higher versus accessing the data consumers. The efficiency of the subsidising mechanism mainly depends on the mutual impact of the number of data consumers and data providers on each others. The mutual impact of the number of data consumers and data providers is refereed to in the thesis by the cross-group externalities term. The data is treated as economic goods by connecting its monetary value with the actual need of the data consumers, which is formalized mathematically by a realistic consumer demand function. As a result of this consideration, i.e., the data as an economic good, the cross-group externalities take mathematically a non-linear form. However, the original theories of the two-sided market model have been studied in the presence of linear cross-group externalities. Consequently, we have revisited the two-sided theories under the non-linear externalities. In addition, we have derived a theory that determines which market side, i.e., data providers and data consumers, has to be subsidized by the data market platform. To the best of our knowledge, we have not seen an explicit condition indicating which market sides the two-sided market model has to subsidize. Moreover, we have studied the question "After the data providers and data consumers meet each others via the data market platform, why data providers and consumers do not negotiate trading the data directly without the platform to get rid of the transaction fees imposed by the platform ?", which is namely by Coase theorem. This question has been skipped in the original theory of the two-sided market theory where it has been assumed that the Coase theorm do not apply to the interactions between the market sides. This contribution is published in [17, 16].

(2) **Contribution 2**: to resolve the subproblem **P2**, we propose a novel game model to reshape the business of the cloud computing to act as a data market platform. The proposed game model integrates the elasticity of the cloud computing with the two-sided market theory in such a way maximizing the profit of all involved parties, i.e., data providers, data consumers and cloud computing, and distributing dynamically the cloud resources among the computational

tasks associated with data processing. The proposed game model comes up with a solution for uncertain externalities in the two sided market model. On other hand, the proposed game model provides clear and efficient mechanism to split the revenues among the involved parties, which is not addressed by the corresponding collaborative models in the literature. This contribution is published in [14].

(3) **Contribution 3**: to resolve the subproblem **P3**, we design a strategic game that aims to deliver complementary data services among multiple data providers over a cloud intermediary platform. More specifically, we formalize the problem as an extended two-sided market model by courting on one side some influential data providers in order to attract other data providers on the same side to form a bundling of data services. The proposed game helps the cloud computing to answer the following questions: 1) Which data providers will be selected to participate in the bundling strategies? 2) Which ones of the selected providers will be subsidized? 3) How the selected data providers will be incentivized to sustain a maximum revenue and prevent undesired and greedy behaviour? Similarly, from the data providers' perspective, the following questions need to be answered: In the presence of other data providers participating in the bundled complementary data, 1) How should a provider price its own data service? 2) How much should the provider pay to the cloud computing versus the computational fees?

At the technical level, we contribute to the two-sided market theory in the following respects: 1) We consider the dependencies among the players on one side (i.e., data providers' side); and 2) we add a two-stage subsidizing: one to attract some data providers, and another one to attract the AI-driven services. On top of this, we embed double cross-group externalities across the players as follows: a) cross-group externalities between data providers and AI-driven services, which are common in two-sided markets; and b) cross-group externalities between the cloud platform and the AI-driven services. The latest one adjusts the power balance among the players (i.e., cloud and data providers) and alleviates the competition between them. In these proposed settings, we derive the new equilibrium of our two-sided market scenario. To the best of our knowledge, our work is the first that capitalizes on the two-sided market theory as a platform to get on board services for complementary data. This

contribution is published in [15].

(4) **Contribution 4**: to resolve the subproblem **P4**, we have proposed a novel double two-sided game that models the interactions among the involved parties (i.e., blockchain node, data providers and data consumers) using the two-sided market theory. In the proposed game, both the data providers and blockchain node act as a two-sided platform that gets on board two market sides. Specifically, the blockchain node intermediates the interactions between the data providers and data consumers, while the data providers intermediate the interactions between the blockchain node and data consumers. The data providers either 1) subsidize the blockchain node by a higher portion of revenue to motivate it to supply more mining computational units, which results in attracting more data consumers and increasing the revenue; or 2) subsidize the data consumers by more data computational units, which increases the consumers' demand and hence contributes in attracting the blockchain node. Similar strategies are set up to the blockchain node. The proposed game combines both strategies as two separate games. The solution of the games helps derive the equilibrium in terms of shared revenue among the blockchain node and data providers and amount of mining resources that each smart contract should be assigned with. This contribution is published in [13].

1.4 Thesis Problem Summary and Organization

We summarize the problem statements and the technical contributions in Figure 1.1. The figure also shows the organization of the thesis by clarifying the content of each chapter.

Research questions	The state of the art solutions	The state of the art drawbacks and research gap	Our proposed solutions	Technical contributions	Thesis chapter
What is the most optimal business model that can be used to design the data market platform ?	Peer-to-Peer model	Not involving actual data owners Raising privacy challenges Violating law regulations associated with selling data	Two-sided market theory	Revisting the theory under non linear externalities	Chapter 2
	Merchant Model	Leading to an aggressive competition among the involved parties! Entailing high operational costs		Revisting the theory under the presence of the Coase theorem	
		Having not studied yet the idea of data market where the data is shown as commercial commodities, and involved parties (i.e data providers and data consumers) find, meet, and matched each other effectively.		Improving the subsidizng mechanisim	
How can the business of the cloud computing be reshaped to act as a data market platform?	Compitive game model	Leading to coarse distribution/provision for cloud resources	Two-sided game model	Developing an elastic game model on the top of the two-sided market	Chapter 3
	Collaborative model	Incapacitating to provide an efficient mechanisms to split earned revenues Not dealing with irrational parties			
How can we design a strategic game that aims to deliver complementary data services among multiple data providers over the two-sided cloud data market platform ?	Bundling collaborative models	Limited for small number of collaborative parties	Two-sided cooperative game model	Developing a two stages subsidizing game model to attract most influential data providers, and another one to attract the AI-driven services.	Chapter 4
How can the blockchain act as data market platform? How can we design a strategic game that make abalance between delivering high level of security services and trusted data transacttions on time from a hand, and the cost of the mining computational units ?		there is lack of attention on the business model that would enable data trading over blockchain. In particular, assigning optimal amount of computational units to the mining tasks, sustaining optimal payoffs to involved players and serving data requests on time	Douple two-sided game model	Devloping double two-sided game that models the interactions among two paltforms acting as a two-sided market.	Chapter 5

Figure 1.1: Problem and Contribution Summary

Chapter 2

Toward Monetizing Personal Data: A Two-Sided Market Analysis

With the increasing popularity of social mobile applications and mobile crowd sensing, holders of smart devices are generating a huge amount of personal data. Nowadays, a wide variety of domains ranging from health-care applications to pollution monitoring are benefiting from collected data. In fact, these personal data may have a monetary value and currently, secondary data owners (such as clinics, Facebook and Twitter) are getting benefit from them either by reselling these data to third entities or by generating statistical analysis. Unfortunately, the primary data owners, the users themselves, are not getting benefit from these transactions. Today, there is no platform to help users monetize their own personal data. In this paper, we propose a two-sided market-based platform for monetizing personal data. Given the intrinsic properties of data as economic good, we prove formally that two-sided market is a realistic solution as it can offer the service of collecting the required data amount and within the quality range required by the buyers. More precisely, 1) we study the two-sided platform equilibrium under non-linear externalities and extract mathematically the condition that states which side will be subsidized by the platform; 2) we study formally the impact of the direct sale on the platform payoff and show that the platform payoff is given by a logarithmic function of end users stability in the platform; and finally 3) using a real data set from Amazon, we construct an empirical comparison between the two-sided platform model and the classic merchant model. In addition, we simulate the efficiency of the two-sided market model in presence of the direct sale. Simulation results show that our two-sided market platform can play a

critical role in motivating users to share their personal data and can be a practical solution for data generated from mobile crowd sensing.

2.1 Introduction

With the emergence of numerous advanced computing services deployed over the cloud, a massive amount of raw data is collected and stored in cloud data centers [56]. According to Cissco, 600 ZB per year will be generated by IOT devices and smart phone users by 2020 [46]. Facebook, as another example, daily captures 900 million of social transactions including uploading photos, videos and text messages performed by 1.13 billion users around the world [56]. This explosive increase of the amount of data, known as big data paradigm, creates further values for wide variety of domains ranging from health-care applications to population monitoring that benefit from such collected data. Market analysts, for example, collect and combine personal data to study the consumer behavior and determine the future directions and pricing processes for certain products. Retailers, banks and insurance firms are further examples that collect and combine personal data to target their consumers with marketing offers and promotional products. This explosive demand on data resources opens the door for multi-billions businesses involving buying and selling consumer's data.

According to the international data corporation, big data market will reach $203 billion by 2020 [46, 1]. In parallel, research and industry communities pay recently attention toward blockchain technologies that opt viable solutions to generate abundant, secure and complete raw data. As a result of this attention, the data contained in the blockchain ledger is expected to worth up to 20% of the global big data market and generate up to 100 billion in annual income [1]. This exponential growth of the big data market raises the need to develop an economic platform that efficiently monetize the data on the cloud. Given the intrinsic properties of data as economic good, the data market platform connects on board the data providers (sellers) and data consumers (buyers) and enables sharing and trading their data commodities. Designing such a platform entails difficult challenges related to data pricing, modeling interactions between involved parties, and maximizing total surplus for the involved parties.

Problem Statement: Despite the manifest economic value of data and significant need for trading data platforms, little attention has been paid toward the development of these platforms. In fact, the main stream research in the context of data focuses on developing mechanisms of data resource management and modeling AI/machine learning networks operating on top of big data layers to extract further insights. However, only few proposals such as [44, 18, 72, 55, 70, 48, 46] address the problems of data valuation and pricing schemes. Those proposals offer information services, which collect and aggregate personal data from individuals for specialized applications. Nevertheless, the research status is still at its infancy and suffers from significant drawbacks. Specifically, it primarily focuses on the trading of data from the perspective of organizations without involving individuals - the actual data owners. Moreover, the current research works in general have inspired by the merchant model and classical economic approaches in which a third-party buys personal data from their owners, reprocesses them, extracts information and sells it for consumers. However, the diversity of the consumer's interests and the huge amount of data residing on cloud servers play against such a model as a successful platform for data trading. From an economic perspective, the merchant model entails higher processing costs when one wants to achieve higher product variety. Practically, the high diversity of consumers adds higher technical data processing costs to produce different types of information, which affects negatively the quality of generated information.

The data market is classified into two categories, namely peer-to-peer model and merchant model. Under the peer-to-peer model, proposals such as [52, 44, 18] focus and discuss the relationship between organizations (i.e., second owners) and consumers without involving the actual data owners. Specifically, organizations (second data owners) benefit from the data of individuals by selling them while the actual owners are not appropriately compensated. Facebook, as a practical example, earned a total revenue of 3.85 billion in the fourth quarter of 2014 from ads [3] - a fact that allows marketers to reach the personal data of their users. Credit bureaux, such as Equifax, Experian and TransUnion sell their consumers' personal data to retailers, banks, insurance firms, and government agencies. The federal agency for Medicare and Medicaid Services, another class of organizations that are engaged in the business of buying and selling personal data, sells medical claims that include medical, demographic, and geographic personnel data to third parties [4]. This

model raises challenging issues related to privacy, and recently selling data to a third party is being seriously discussed by law communities. For example, the General Data Protection Regulation (GDPR)[1] calls for mandating the process of exchanging personal data by law to protect the fundamental right of actual data owners and allow them to have a control on their personal data. Such a call requires a solid and more coherent monetizing data framework backed by a strong market model to incorporate the primary data owners in the process of buying and selling data. Specifically, the monetizing data framework has to develop an effective mechanism to guarantee an adequate participation of these owners in the trading process.

Under the merchant model [11], monopolistic organizations take the possession of seller's data, reprocess them, and resell them to consumers at retrial price. In fact, those organizations act as information service providers rather than an open platform allowing data consumers to access the huge pool of raw data. Such a model (merchant model) for monetizing big amounts of data over the cloud raises serious challenges related to revenue maximization and leads to an ineffective third party trading schema. Specifically, studies adopting the merchant model such as [70, 48] and [46] overlook its limitations that include high processing/operation costs once a merchant wants to achieve higher product verity in terms of type, quality and quantity of products. Moreover, under the merchant model, the data providers (actual data owners) aim to maximize their revenue from their data commodities while the cloud provider (information service provider as third party) aims to minimize the cost of the raw data. In parallel, at the information consumer-side, the cloud aims to maximize its revenue from selling the processed information while the consumer aims to minimize the cost of information commodities considering the maximum available quality and quantity of information. The resulting equilibrium from such aggressive competitions among the involved parties leads to less and coarse distribution for total surplus. In addition, data differ from other economic goods for the possibility to be resold to many consumers at the same time. In the merchant model, because of selling economic goods that cannot be resold, the equilibrium of the market is given by the intersection of the demand and supply curves, which means the quantities of goods needed by the consumers equal the quantities of goods provided by the sellers. This does not hold in our case since the same data can be shared between all the consumers. For instance, if there are 1000 consumers

[1] https://gdpr-info.eu/

and each consumer requires 100 units of particular data, it does not mean that 100000 data units are needed, where it is enough for the merchant to buy 100 units and share them with the whole data consumers. This will largely raise the competition between data providers and push them to accept lower prices, which negatively affects their total surpluses and leads to inadequate involving of providers in the process of monetizing data.

Recently, commercial platforms such as People.io[2], Opiria[3], and Lotame[4] have been launched to provide solutions for personal data trading. Those platforms deliver social channels between data providers and data consumers and get them on board. Specifically, the platforms grant a partial control for people and allow them to monetize their data by licensing them to brands in return for a payment/reward. Moreover, the platforms offer a grounded basis toward user's data privacy wherein people are able to see when, where and how their data is used. However, these platforms, generally initiated for a commercial purpose, are relatively new, and their business models and theoretical foundations are not revealed. Their efficiencies and limitations have not been analyzed and studied yet. Motivated by the promising insights seen from industrial lens, we revise and extend our previous work [17] that introduced the two-sided data monetization model. To the best of our knowledge, this paper is the first scientific initiative that 1) provides a comprehensive study for the two-sided model [84] as a solution and business model for data trading intermediaries and 2) contributes to the theoretical and technical details of data monetization platforms.

2.1.1 Motivations and Contributions

To summarize, the data market is facing the following problems:

(1) There is a lack of platform for monetizing data that involves a wide range of primary data providers. In fact, secondary data owners, such as social media providers, usually control this operation and get the full benefit from it.

(2) The merchant and peer-to-peer models are not suitable as business models for the current data market. Individuals, i.e., primary data owners, are either not compensated for sharing their

[2] https://econsultancy.com/start-me-up-people-io-allows-people-to-monetize-their-personal-data/
[3] https://medium.com/@EVALUAPE1/blockchain-project-review-opiria-6-3-dapp-for-data-exchange-fdc78a8be7b7
[4] https://www.lotame.com/its-time-to-unstack/

personal data or compensated by non-monetary rewards; and the wide range of inhomogeneous data consumers with different interests entitles high operational costs on the merchant model.

(3) Recently, some intermediary commercial platforms to trade personal data have been lunched. However, only white papers describing these platforms from a high-level perspective are available without disclosing and revealing the grounded business model and its internal economical foundations and details. There is a lack of research proposals that formalize such platforms, try to provide insight into operational mechanisms, and analyze their efficiencies.

To mitigate the aforementioned problems, we propose a business model using the two-sided market theory [84], described in details in Section 2.3, for a data monetization platform. The proposed model consists of three entities: data providers, data consumers and data intermediary platform. The intermediary platform acts as a mall store in which the providers and consumers meet to sell/buy data. The data providers and consumers (market sides) exhibit cross-group network effects (externalities- described in details in Section 2.3). The data platform acquires revenues by imposing transaction fees on market sides. The revenue is maximized using a subsidy technique. The subsidy technique, described in details in Section 2.3, attracts one of the market sides by charging low or even no transaction fees, while the other market side is enticed by the demand size of the subsidized one. We contribute to the literature of data market as follows:

(1) We propose a new business model for data monetization platforms. A significant characteristic of the model is that it moves the control to individuals (i.e., primary data owners) and enables them to offer their personal data as economic goods, which means aligning international and law regulations that focus on the privacy of data providers. Moreover, the model helps different types of data consumers get a quick and easy access to large pools of high quality data by increasing the engagement of data providers using theory of subsidizing market sides of the two-sided model.

(2) We introduce a mathematical model that enables increasing the total surplus for the data providers, data consumers and platform, which outweigh the total surplus resulted by the classical merchant model.

(3) We investigate the efficiency and detect the potential limitations of some recent data monetization platforms launched for commercial purposes.

At the technical level, we contribute to the two-sided market literature as follows:

(1) We revisit the two-sided model's equilibrium under non-linear externalities market sides. Specifically;

- We re-identify the original price structure of the two-sided market model introduced by [82].
- We derive the condition that determines which market side has to be subsidized by the platform. To the best of our knowledge, we have not seen an explicit condition indicating which market sides the platform has to subsidize.
- We re-investigate the equilibrium of the two-sided model over different ranges of externalities. We show that Assumption 1 (the well-known assumption on the two-sided market model stating that externalities are not too strong) is not enough to conclude about the efficiency of the two-sided model.

(2) We investigate the two-sided model's equilibrium in presence of the Coase theorem [28] (explained in details in Section 2.3). Specifically, we allow the market sides to negotiate with each other to bypass the platform and trade directly.

(3) We conduct a simulation analysis to compare the efficiency of the two-sided market model with the classical broker form (merchant model).

2.2 Research Methodology and Paper Overview

In this section, we describe step by step the research methodology followed in this paper. As shown in Figure 2.1, the paper structure and research methodology are divided into four main stages: 1) Data monetization platform; 2) Direct sale impact on the two-sided data monetization platform; 3) Simulation and 4) Discussion. Each stage is divided into steps as follows:

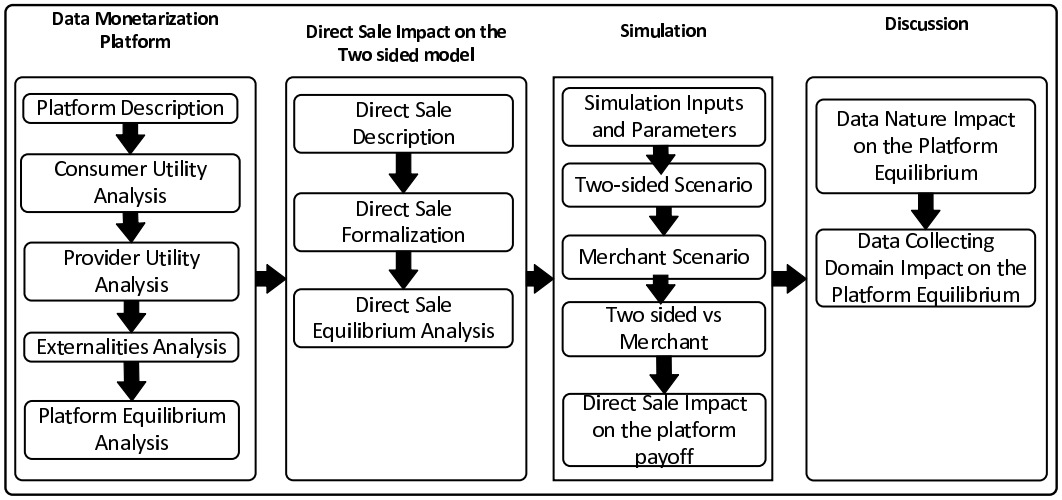


Figure 2.1: Paper overview and research methodology

- Data monetization platform: under this stage, we design a two-sided market platform for data trading. This stage is divided into the following steps:

  i. Platform description: In this step, the components of the proposed model and how it works are described. This step is addressed in Section 2.4.1.

  ii. Consumer utility analysis: In this step, the requirements and potential behaviors of data consumers are analyzed and discussed. Principally, we discuss the valuation of data with respect to the consumers. Thereafter, we choose our parameters and formalize the utilities of consumers as given in Equation 3. This step is addressed in Section 2.4.2.

  iii. Provider utility analysis: In this step, the potential behaviors of data providers are analyzed and discussed. Thereafter, we choose our parameters and formalize the utilities of providers as given in Equation 4. This step is addressed in Section 2.4.3.

  iv. Externalities analysis: In this step, we characterize the proper form of the mathematical function that represents demand curves and externalities between the market sides. Specifically, we use the consumer and provider analysis constructed into two previous

steps to represent demand curves. The constructed analysis leads to formalize the demand curves using a mathematical logarithmic function as given in Equations 5 and 6. This step is addressed in Section 2.4.4.

v. Platform equilibrium analysis: In this step, we identify the structure of platform optimal fees. This sub-step is addressed in Section 34. Thereafter, the platform equilibrium is extensively investigated and given as a function of externalities elasticity. Specifically, we highlight the efficiency of the platform over different ranges of externalities. This step is addressed in Section 2.4.6.

- Direct sale impact on the two-sided model: In this stage, we study the effects of direct interactions between consumers and providers on the two-sided platform efficiency. Specifically, this stage aims to discuss the effect of the Coase theorem (explained in details in Section 2.3) on the two-sided data platform. This stage is divided into the following steps:

  i. Direct sale description: In this step, we describe the case when the market sides negotiate with each other to bypass the platform. This step is addressed in Section 2.5.1.

  ii. Direct sale formalization: In this step, we formalize the interactions described in the previous step as a game established among the consumers, providers and platform. The formalization aims to analyze the platform efficiency in presence of the direct sale. This step is addressed in Section 2.5.2.

  iii. Direct sale equilibrium analysis: In this step, we extensively investigate the equilibrium of the game formalized in the previous step. In the investigation, we essentially focus on 1) the possibility of the success of the Coase theorem at long run time, i.e., following many transactions between data providers and data consumers; and 2) if the Coase theorem takes place, to which extent this will affect the efficiency of the platform. We find that the platform efficiency depends on the mobility/stability of data providers in the platform. This step is addressed in Section 2.5.3.

- Simulation: In this stage, we implement a case study using a real dataset to validate the two-sided model as a solution for data monetization. This stage is divided into the following

steps:

(1) Simulation inputs and parameters: In this step, we describe in details the inputs and adjustable parameters of the simulation. Specifically, a real dataset is introduced and described including the distributions of demand and utility functions. This step is addressed in Section 2.6.1.

(2) Two-sided market scenario: In this step, we describe the implemented two-sided market scenario. This step is addressed in Section 2.6.2.

(3) Merchant scenario: In this step, we describe the implemented merchant model scenario. This step is addressed in Section 2.6.3.

(4) Two-sided vs merchant: In this step, we evaluate the efficiency of both the two-sided market and merchant models. The simulation results show that the proposed platform outperforms the classical intermediaries model (merchant model) in terms of profit. This step is addressed in Section 2.6.4.

(5) Direct sale impact on the platform payoff: In this step, we demonstrate the unregulated case using a simulation scenario. The proposed platform shows higher efficiency in domains in which providers exhibit more unstable mobility such as mobile phone sensing networks. This step is addressed in Section 2.6.5

- Discussion: In this stage, we discuss the theoretical and simulation results in terms of data market. Specifically, we discuss the following two questions:

(1) "How does the data nature affect the equilibrium of the two-sided market platform?": Under this question, we discuss the key aspects for data as an economic good, and connect those aspects to theoretical results obtained in Section 2.4.6. The discussion aims to highlight the platform behavior under different data types and circumstances. This question is addressed in Section 2.7.1.

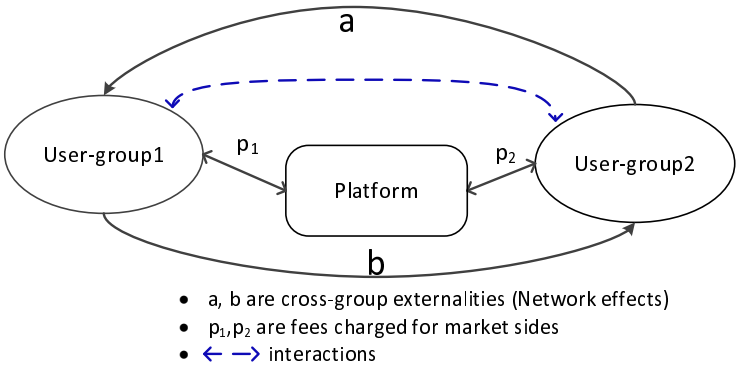


Figure 2.2: Two-sided market platform

(2) "How does the domain of collecting and sharing data affect the equilibrium of the two-sided market platform?": Under this question, we connect different domains of collecting and sharing data to the Coase theorem, and then we discuss the platform effectiveness in presence of the theoretical and simulation results presented in Sections 2.5.3 and 2.6.5 respectively. The discussion highlights the domains in which the two-sided market shows promising efficiency. This question is addressed in Section 2.7.2.

2.3 Background:Two-Sided Market

This section provides a general overview of seminal two-sided market papers [21, 82, 83, 25, 84, 12, 11]. The overview scope is narrowed down to focus on the perspectives employed in this study, putting aside the advanced economics-related details that are not relevant to this study.

Two-sided market is defined as a market where one or several platforms enable interactions between two distinct end-user groups that can exhibit mutual benefits or perform commercial transactions. The platform tries to connect these groups on board and charges them appropriately with fees. As shown in Figure 5.3, the platform connects *User-group 1* and *User-group 2* on board, where they start their interactions. For example, credit cards such as Visa, Mastercard and American Express, are two-sided platforms where: 1) the card through its financial organization presents the platform; 2) the credit card holders present *User-group 1*; and 3) the merchants/stores present *User-group 2*.

Figure 5.3 shows a loop of network effects a and b, called *cross-group externalities*, between

User-group 1 and *User-group 2*. Cross-group externalities represent the benefits that market sides obtain from increasing their participation. In other words, the externality a represents the amount of benefits a user from *User-group 1* (e.g., providers) obtains when one more user is added to *User-group 2* (e.g., consumers). Similarly, the externality b represents the amount benefits a user from *User-group 2* obtains when one more user is added to *User-group 1*. When these benefits can be expressed using linear equations, the externality is said to be linear. Otherwise, the externality is non-linear. An example of non-linear externalities is when logarithmic functions are used to express the network effect. The existence of externalities is the first characteristic that distinguishes two-sided market from other types of platforms. In the credit card example, the externalities between card holders and merchants are as follows: merchants pick the platforms (i.e., the cards) that are or will be popular among consumers. At the same time, consumers (i.e., the cardholders) pick the cards that are or will be accepted by most merchants.

The market sides form a loop of externalities that is defined as follows: the users of *User-group 1* are incited by the size of *User-group 2* and the users of *User-group 2* are incited by the size of *User-group 1* at the same time. Becuase of this loop of externalities, the only way to connect both market sides on board is by attracting user groups through incentives, rather than through the size of the other side, for instance by charging them with low or no fees. This entails that the two-sided platform discriminates the fees between its user groups. Figure 5.3 shows that the platform charges the market sides with different fees p_1 and p_2. This technique is called the *subsidy technique* and the attracted side is called the *the subsidized side*. Once the subsidized side connects to the platform, the other side (i.e., the subsidizing side) is attracted by the size of the subsidized side and connects to the platform. In the credit card example, credit companies need to attract holders in order to convince merchants to accept their cards and they need merchants accepting their cards to induce consumers to use these cards. Credit card companies attract first the consumers by offering them their cards for free (i.e., zero transaction fee p_1) or low annual fee, while they make benefits on the merchants side through per-transaction fee (i.e., $p_2 > 0$).

User groups can be charged in many forms and over many stages such as fixed affiliation fees and per-transaction fees. The charging form depends primarily on the nature of the business of the market sides and the strength of the externalities between them. In case the externalities are

relatively weak, the platform uses the affiliation fees since the user's payment does not explicitly depend on how the platform performs on the other side of the market. By contrast, the platform chooses the per-transaction mode in case the externalities are relatively strong. In this case, user payment might be an explicit function of the platform performance on the other side. However, if a participant pays to a platform only in the event of a successful transaction, the user does not need to worry about how the platform does in dealing with the other side [12].

In two-sided market, users can choose to use one or many platforms. When users choose to use only one platform, the users are commonly said to be "single-homing". When users use several platforms, they are said to be "multi-homing". For example, credit card holders often choose one type of cards, for example Visa, while merchants may accept dealing with many cards such as Mastercard and Visa. In this case, card holders are single-home while merchants are multi-home.

Definition 1. *[83] The platform is a two-sided market platform if*

(1) The Coase theorem [28] does not apply to the relationship between market sides

(2) The volume of transaction performed by market sides depends on the structure fees level, not on the total fees level

The first term of the two-sided market definition (i.e., the Coase theorem) states that the market sides (i.e., User-group 1 and User-group 2 in Figure 2) cannot negotiate with each other to bypass the platform and trade directly. Let us consider the credit card example again, assume that all the merchants are 1) willing to accept a non cash/debit payment from their buyers with the same or less fees than the credit card's fees; and 2) able to bear the risk that some buyers are not fulfilling their commitments (i.e., not paying their bills). In such a case, merchants can negotiate and successfully convince the holders to trade directly and use their offers instead of credit cards.

The second term of the two-sided market definition refers implicitly to the cross-group externalities. Specifically, changing the charging fees structure affects the number of attracted users, and hence affects the total transactions among market sides. For example, let us assume that the credit card company charges the merchants with 2% for each transaction paid using the credit card, while the customers (i.e., the card holders) do not bear any fees. If the credit card company increases the fees on the holders side and reduces the percentage paid by the merchants by an equal amount for

instance (i.e., holders and merchants are equally charged 1% each for every transaction), the number of the holders will decrease and hence the total transactions will decrease as well.

It is important to mention that the main difference between the classical intermediates and the two-sided market is that the classical intermediates acquire goods at a whole price from sellers and resell them to buyers at a retail price. Sellers' and buyers' decision to involve in the selling process depends only on price offered by those intermediates. Walmart and Amazon are two examples of classical intermediates. By contrast, in the two-sided market, the seller and buyer decision depends on fees offered by the platform and the number of affiliated other sellers and buyers.

Finally, the total joint surplus refers to the total of the buyer's surplus and the seller's surplus. The buyer's surplus occurs when the price for a product or service is lower than the highest price the buyer is willing to pay. Joint surplus is used to describe the equilibrium of the market and measures the satisfaction of the consumers and sellers. We will use this concept to prove that consumers and providers perform better in the presence of the direct sale, i.e. they receive more surplus.

2.4 Platform Model

This section provides the technical details of the two-sided data monetization platform. Mainly, the section starts by a general description of the proposed model including the involved parties and their interactions in Section 2.4.1. Thereafter, the strategies of these parties including their utility and demand functions are formalized in Sections 2.4.2, 2.4.3,2.4.4, and 34. Finally, the platform's equilibrium is formalized and investigated in Section 2.4.6.

2.4.1 Platform Description

The proposed platform, as shown in Figure 2.3, consists of three parties: *Data Providers*, *Data Consumers* and a *Data Broker*. Data consumers are organizations whose business often requires huge amount of data with particular specification to perform some business related analysis. Consumers' requirements vary in terms of the type, quality and amount of data based on their scope and their applications' needs. Providers can be smartphone users, individuals, or professional IOT sensor providers having some personal data to sell. We assume that all data providers behave rationally and are willing to sell their data at market retrial price p. The broker is an online or cloud

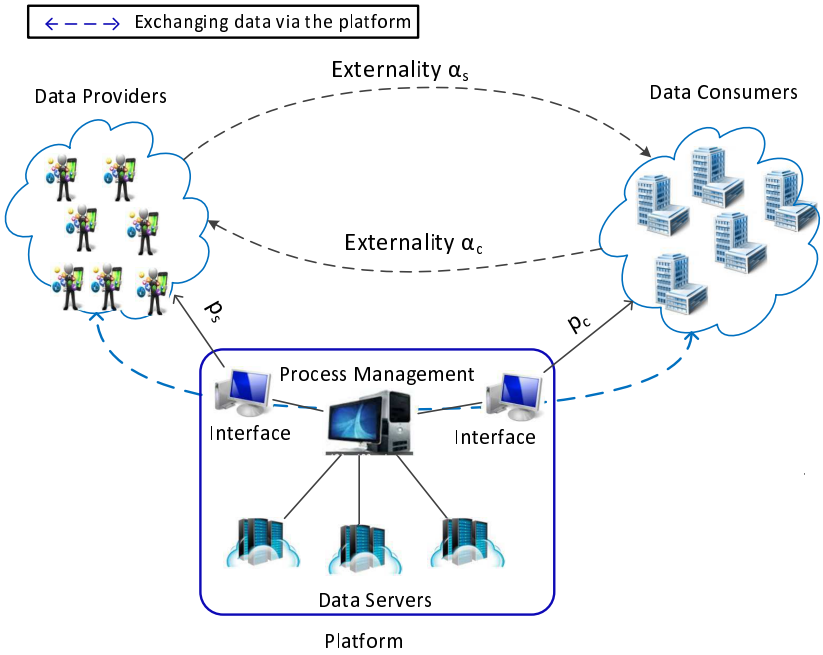


Figure 2.3: Personal data monetization platform

platform equipped with the needed infrastructure to store and share data. The broker provides services that enable data providers and data consumers to perform data selling and buying transactions. Once a consumer sends a request to the broker, the broker matches providers meeting consumer's requirements and connect them on board. Consumers and providers interact and negotiate through communication channels provided by the platform (broker). The broker charges providers and consumers with transaction fees of p_s and p_c respectively. The mathematical notations used in this paper are summarized in Table 2.1.

The price of data (p) is pre-defined by its providers (i.e., posted price). In fact, there are two main categories of sale mechanisms: posted prices and negotiation/auctions deals. Under the posted prices category, sellers have some expectations about market status (i.e., demand and competition) and set their prices accordingly. Amazon, for example, is a two-sided platform that uses the posted prices form [93]. Auctions can be used as a price discovery mechanism in the case of imperfect information about the market status (i.e., the actual value of a commercial good is not known for buyers or sellers). However, auctions are limited for relatively small size of demands and non-resold commercial goods, which is not applicable for our scope (monetizing personal data). Specifically, the same data can be shared between many data consumers. Thus, in our platform, the data provider

maximizes the payoff by selling the data to a maximum number of consumers, while the same provider maximizes the payoff by maximizing the selling data price for one data consumer in the auction form. Furthermore, at the level of big data trading, consumers buy relatively big chunks of data that are combined from many providers. Thus, reverse auction in which sellers bid for the prices, is not applicable as well for the case of monetizing personal data. In fact, consumers are not buying their data from one seller, and thus, as argued in [93], reverse auction is useless in such cases.

The broker chooses the per-transaction fees mode because consumers may ask for different amounts of data. Thus, the platform cannot impose the same affiliation fees on consumers who require a relatively small data amount and consumers who require a relatively big data amount. Choosing the form of fees (per-transaction or affiliation) is a problem of maximizing the broker's payoff. The per-transaction mode is similar when companies offer different plans to their customers. Customers with less ability to pay choose a lower plan while customers with higher ability choose a higher plan.

Assumptions: we assume that all data providers, data consumers and the platform behave rationally and are willing to enroll in the business of data market. The rational behavior refers to the ability of providers, consumers, and the platform to build their decisions based on their preferences and according to the available information about the data market. That is, data providers are aware of the market retrial price p for their data. Data consumers have clear preferences and are aware of the set of choices and alternatives in terms of the desired amount and quality of data, and they take their decisions considering their budgets. The platform is aware of the status of the data market including consumers demand and available data supplying. In fact, the platform is supported by required techniques to 1) observe and measure key factors affecting the data market, such as the willingness of consumers to buy the data, their own budgets, and size of their business; 2) deal with uncertainty and imperfect information about consumers and providers; and 3) perform some optimization before taking actions. The ultimate goal of data providers, consumers and the platform is to maximize their own benefits according to the available information. Specifically, the platform specifies the charging transaction fees on consumers and providers, data providers decide on their own data price, and data consumers determine the size and quality of data to be bought. Another

assumption made in the paper is about imitating the real market, in which there is no monopoly on either sides of the market and there is relatively large supply and demand.

Table 2.1: Model parameters

Notation	Definition	Notation	Definition
p	Data retrial market price.	p_c	The transaction fee imposed by the platform on the consumers side .
p_s	The transaction fee imposed by the platform on the providers side.	α_c	the externality from consumers to providers.
α_s	The externality from providers to consumers.	$v_i(n\|q)$	The value function that defines the utilities that a consumer received from using n data amount given a certain data quality q.
n	Data amount.	UC_i	The utility that the consumer i receives from interacting over the platform.
US_j	The utility that the provider j receives from interacting over the platform.	$N_{s\|q}$	The number of active providers that are connected to the platform and provide data quality q.
γ	The average amount of data provided and sold by each data provider.	$N_{c\|q}$	The number of active consumers that are connected to the platform and require amount of data with quality q.
β_s	The slop of the providers demand with respect to p_s.	β_c	The slop of the consumers demand with respect to p_c.
ϵ_c	A constant defined in the consumers demand function.	ϵ_s	A constant defined in the providers demand function.
$\eta^c_{p_c}$	The elasticity of consumers demand with respect to the per transaction fee p_c.	$\eta^c_{p_s}$	The elasticity of consumers demand with respect to the per transaction fee p_s.
$\eta^p_{p_c}$	The elasticity of providers demand with respect to the per transaction fee p_c.	$\eta^p_{p_s}$	The elasticity of providers demand with respect to the per transaction fee p_s.
π	Platform's payoff.	f	A fixed cost per transaction between consumers and providers.
λ	The rate of data requests performed by each consumer.	n'	The maximum amount of sufficient data that a consumer receive from the direct sale.
y	The percentage of the data amount that a consumer requires from the maximum data that the platform can offer.	T	total transactions performed between data consumers and providers over the platform.
f_c	The per-transaction cost that a consumer incurs when he decides to interact directly without the platform.	f_s	The per-transaction cost that a provider incurs when he decides to interact directly without the platform.

Continued on next page

Table 2.1 – *Continued from previous page*

Notation	Definition	Notation	Definition
p_i	The per transaction incentive that a consumer/provider offers for providers/consumers to convince them to interact directly.	p_c^*	The Nash per transaction fee imposed by the platform on the consumers sides when it competes with the direct sale.
p_s^*	The Nash per transaction fee imposed by the platform on the providers side when it competes with the direct sale.	π^*	The Nash platform payoff.
r	The number of requests after which a consumer will never interact over the platform.	Δ	The difference between the maximum sufficient amount of data that the a consumer requires and the last maximum sufficient amount of data that the platform is able to provide for the consumer.
π_m	Merchant's payoff.	TS_m	Merchant model's total surplus.
x	A random number that represents the probability of each provider to 1) perform the process of selling/buying data and 2) meet the requirement of a new consumer request.		α_c, β_c, ϵ_c, α_s, β_s, and ϵ_s are positives

2.4.2 Data Consumers

Data consumers are organizations whose business often requires a certain amount of data with particular specification. It is important to mention that consumers can range from individuals to big organizations. In this work, we consider as a running example the case of the consumers in mobile sensing domain such as [22, 20, 63, 51, 29, 67]. These consumers collect data from mobile phone users to perform certain sensing services that some times require an installation for applications on the users' mobile devices.

In this example, two issues are to be considered: (i) The inadequate level of participation in the sensing services due to the absence of incentives in the volunteering-based action; and (ii) the high costs of incentive-based action to guarantee an adequate level of participation, which outweighs the consumers' budget and sometimes the real economic value of the required data. The consumer determines an acceptable range of qualities and amounts, out of which the data become useless and the consumer becomes unwilling to purchase them. In addition, data consumers are heterogeneous and vary in their applications' needs. In fact, what is considered as valuable and worthy data for a consumer is considered at the same time as unvaluable and unworthy for another one. This imposes on the platform a high performance (reducing more the transaction fees) to attract data providers side

to meet the consumers' requirements. In terms of the two-sided market, the platform highly subsidizes data providers by charging them a low-transaction fee, reaching zero. Usually, credit cards, as example, do not charge card holders (i.e., zero transaction fees), or charge them a small annual fee as a counter-part of monetary and/or non-monetary advantages (for instance travel rewards), while they charge the merchants.

We define the value function $v_i(n|q)$ that specifies the economic value for the consumer i. $v_i(n|q)$ measures the benefit of using a certain amount (n) from the data quality q provided by the platform. The value function in general represents the monetary benefit that a commodity provides to the buyers. This is known in Economics by "value in use (Adam Smith Theory of Value)" [90]. In our context, the value function denotes how much value the data add to the buyer, which can be represented by how much the buyer is willing to pay in order to get this kind of data. For example, commercial firms such as retailers, banks and insurance companies collect and combine personal data including their consumers contacts to target them with marketing offers and promotional products. Let us assume that the net payoff of each product is 100 USD, 1000 consumers are targeted and 30 percent of them are buying those offered products. Then, the value of data reaches up to 30000 USD. Two factors affect $v_i(n|q)$: data quality provided by the platform and the amount of data. Each consumer has a minimum amount of data (min_n). Under this amount, the data becomes inefficient for the consumer. This means that the value function $v_i(n|q)$ becomes equal to zero when the amount is less than the minimum amount min_n and the consumer becomes unwilling to pay. After min_n, the value function $v_i(n|q)$ increases at a different rate. However, the value function does not increase linearly with respect to the amount of data. Moreover, after a certain amount, the value function stabilizes and the consumer becomes unwilling to purchase more. These observations are well-grounded in prior analytical and empirical studies in the same or a similar context [68, 66, 54, 55]. Based on these observations, the function $v_i(n|q)$ should satisfy the following properties:

(1) For each consumer i, there is acceptable range of data amounts n defined by the continuous interval $[min_n, max_n]$.

(2) The value function $v_i(n|q) = 0$ when $n < min_n$

(3) The value function $v_i(n|q) = Max(v)$ when $n > max_n$

(4) The value function $v_i(n|q)$ is a monotonically increasing concave function when $n \in [min_n, max_n]$, thus $\partial v_i(n|q) \setminus \partial n$ equals to zero when $n < min_n$ and $n > max_n$, and the second derivative $\partial v'(n) \setminus \partial n \geq 0$. In terms of two-sided market, the externalities from the consumers' side to the providers side exist only in $[min_n, max_n]$ and vanishes after max_n. This means that the platform highly performs on the providers' side up to a certain threshold of the size of the data providers.

$v_i(n|q)$ could be defined as follows where c_1 is a constant:

$$v_i(n|q) = \begin{cases} a\log(n) - c_1 & \text{if } n \in [min_n, max_n] \\ 0 & \text{if } n < min_n \\ a\log(max_n) - c_1 & \text{if } n > max_n \end{cases} \tag{1}$$

In the platform, data providers provide and sell, in average, γ data units. Thus, we can rewrite the utility of the consumers $v_i(n|q)$ as a function of the number of data providers $N_{s|q}$ whose data quality q as follows:

$$v_i(n|q) = \begin{cases} a\log(N_{s|q}) - b & \text{if } \gamma N_{s|q} \in [min_n, max_n] \\ 0 & \text{if } \gamma N_{s|q} < min_n \\ a\log(max_n) - b & \text{if } \gamma N_{s|q} > max_n \end{cases} \tag{2}$$

where b is a constant and $b = a\log\gamma - c_1$. The platform will use the pattern of the per-transaction fees. The platform matches the consumer's request and charges him with per-data unit fee p_c. The utility of the consumer i is defined as follows:

$$UC_i = \begin{cases} v_i(n|q) - (p + p_c)n & , n \in [min_n, max_n] \\ 0 & , n < min_n \\ v(max_n) - (p + p_c)max_n & , n > max_n \end{cases} \tag{3}$$

2.4.3 Data Providers

Data providers may range from unprofessional individuals in selling their personal data to organizations that provide data. Under the two-sided market model, the platform combines data consumers in one place and incites providers by huge demands, which in turn maximizes their revenues. However, this means that there are externalities from the providers side to the consumers' side and this imposes a certain performance from the platform on the consumers' side. The utility of the provider US_j is given by Equation 4 where $N_{c|q}$ is the consumer demand on the data quality q. It is easy to see that the utility function is monotonic, increasing and linear with respect to the number of consumers and monotonic, decreasing and linear with respect to the per-transaction fees.

$$US_j = \gamma(p - p_s)N_{c|q} \tag{4}$$

2.4.4 Demands

The supply demand is given as function of imposed transaction fee p_s and the expected number of consumers. Similarly, the consumers demand depends on the imposed transaction fee p_c and the expected number of connected providers. As clarified earlier in Section 2.4.2, specifically in the characteristics of the value function given by Equation 1, the value derived from the amount n of data is strictly concave down over n up to a certain threshold of data amount and after this threshold, the value becomes fixed. This means that the consumers' demands are affected by a specific supplying range and the demand is defined as a constant outside this range. In terms of the two-sided market, consumers are not always attracted by more providers. Thus, we use the logit-type demand function introduced in [86] to model the two-sided demands. The logit-type demand function, the most functional form in empirical work, tends to satisfy these characteristics for realistic values of the respective parameters. Consumers and supplying demands are given by Equation 5 and Equation 6 respectively.

$$logN_{c|q} = \alpha_c logN_{s|q} - \beta_c logp_c + \epsilon_c \tag{5}$$

$$logN_{s|q} = \alpha_s logN_{c|q} - \beta_s logp_s + \epsilon_s \tag{6}$$

where α_c is the marginal value that a consumer places on the additional data provider on the platform. Similarly, α_s is the marginal value that a provider places on an additional data consumer on the platform. α_c, α_s represent externalities (network effects) between consumers and providers. β_c, β_s are slopes of consumers and supply demands with respect to transaction fees p_c,p_s respectively. ϵ_c, ϵ_s are constants. For convenience in later calculations, we express demands as functions of transaction fees only as follows.

$$logN_{c|q} = \frac{\beta_c logp_c + \alpha_c \beta_s logp_s - (\alpha_c \epsilon_s + \epsilon_c)}{\alpha_c \alpha_s - 1} \tag{7}$$

$$logN_{s|q} = \frac{\beta_s logp_s + \alpha_s \beta_c logp_c - (\alpha_s \epsilon_c + \epsilon_s)}{\alpha_c \alpha_s - 1} \tag{8}$$

As noted in demand equations, the demand has an inverse relationship with the fees charged by the platform; the demand increases when the platform decreases the charged fees. The effect of fees on the demand is defined by demand elasticity. Demand elasticity is calculated by the percentage of change in the data quantity, divided by the percentage of change in fees. For example, the elasticity of consumers' demand with respect to charged fee p_c is defined by $(\partial N_{c|q}/\partial p_c)$. A higher demand elasticity for fees means that consumers/providers are more responsive to changes in these fees. If the demand for a particular good is more elastic in response to changes in other factors, companies must be more cautious with raising prices of their goods. The percentage of change in the consumers and providers demand (or elasticities of quasi-demands as introduced in [82]) with respect to charged fees p_c and p_s, denoted by $\eta^c_{p_c}, \eta^c_{p_s}, \eta^p_{p_c}$ and $\eta^p_{p_s}$ respectively, are as follows:

$$\begin{aligned} \eta^c_{p_c} &= -\frac{p_c(\partial N_{c|q}/\partial p_c)}{N_{c|q}}, \quad \eta^p_{p_c} = -\frac{p_c(\partial N_{s|q}/\partial p_c)}{N_{s|q}}, \\ \eta^c_{p_s} &= -\frac{p_s(\partial N_{c|q}/\partial p_s)}{N_{c|q}}, \quad \eta^p_{p_s} = -\frac{p_s(\partial N_{s|q}/\partial p_s)}{N_{s|q}} \end{aligned} \tag{9}$$

Similarly, we define the elasticity of total transactions between the consumers and providers $\eta^T_{p_t}$ with respect to the total fees $p_t = p_c + p_s$ as follows:

$$\eta^{T}_{p_t} = -\frac{p_t(\partial T/\partial p_t)}{T} = -\frac{p_t(\partial(\gamma N_{c|q}N_{s|q})/\partial p_t)}{\gamma N_{c|q}N_{s|q}} \\ = -\frac{p_t((\partial N_{c|q}/\partial p_t)N_{s|q} + (\partial N_{s|q}/\partial p_t)N_{c|q})}{N_{c|q}N_{s|q}} \tag{10}$$

where $T = \gamma N_{s|q}N_{c|q}$ is the total transactions of exchanged data from providers to the consumers over the platform.

2.4.5 Monopoly Platform Optimum

Consider the monopoly platform private optimum under which the platform is free to set per-transaction fees p_c and p_s for both data consumers and data providers. The platform faces the problem of choosing p_c and p_s to maximize its payoff which is given by Equation 11:

$$\pi = \gamma(p_c + p_s - f)N_{c|q}N_{s|q} \tag{11}$$

where π is the profit of the platform and f is a fixed cost per-transaction that the platform incurs to connect providers and consumers and to distribute the data from providers to consumers. Because the market sides (consumers and providers sides) provide externalities between each other, the platform faces an inverse relationship between p_c and p_s; that is, maximizing with respect to p_c results in a smaller p_c when p_s is larger. Similarly, maximizing with respect to p_s results in a smaller p_s when p_c is larger. In particular, the optimal p_c for the platform given p_s, defined by $\frac{\partial \pi}{\partial p_c} = 0$, is given as follows:

$$p_c = f - p_s - \frac{N_{c|q}N_{s|q}}{N_{s|q}(\partial N_{c|q}/\partial p_c) + N_{c|q}(\partial N_{s|q}/\partial p_c)} \tag{12}$$

and the optimal p_s for the platform given p_c, defined by $\frac{\partial \pi}{\partial p_s} = 0$, is given by

$$p_s = f - p_c - \frac{N_{c|q}N_{s|q}}{N_{s|q}(\partial N_{c|q}/\partial p_s) + N_{c|q}(\partial N_{s|q}/\partial p_s)} \tag{13}$$

Solving the two equations above represents the condition, defined in Equation 14, characterizing the values of p_c and p_s that maximize the platform profit. This condition shows that the impact of a small (absolute) variation of fees has to be the same on both sides.

$$(\partial N_{c|q}/\partial p_s - \partial N_{c|q}/\partial p_c)N_{s|q} = (\partial N_{s|q}/\partial p_c - \partial N_{s|q}/\partial p_s)N_{c|q} \tag{14}$$

Proposition 1. *i. The fees structure is given by ration of elasticities:*

$$\frac{p_c}{(1+\alpha_s)\eta^c_{p_c}} = \frac{p_s}{(1+\alpha_c)\eta^p_{p_s}} = \frac{\alpha_s p_c}{(1+\alpha_s)\eta^p_{p_c}} = \frac{\alpha_c p_s}{(1+\alpha_c)\eta^c_{p_s}}$$

ii. The elasticity of total transactions between the consumers and providers $\eta^T_{p_t}$ with respect to the total fees $p_t = p_c + p_s$ is given by:

$$\eta^T_{p_t} = \eta^c_{p_c} + \eta^c_{p_s} + \eta^p_{p_c} + \eta^p_{p_s}$$

Proof. See A.1 □

Proposition 1 describes the optimal fees structure that the platform will choose to maximize its payoff. As shown in proposition 1, fees structure depends on externalities between market sides and elasticities of the consumers' and providers' demands. The second part of the proposition 1 states that the volume of transactions depends on the whole elasticity of demands to fees charged by the platform. The second part implicitly suggests that the platform is a two-sided market where elasticities of the volume of transactions are sensitive to the fees structure. Any changes on fees will lead to changes on elasticities of demand curves, which in turn affects the elasticities of the volume of transactions.

2.4.6 Equilibrium Analysis of the Monopoly Platform

The platform relies on the externalities between market sides to create its business. On our platform, the service (i.e. providing data) becomes more valuable for the consumer when more providers become active on the platform and vice verse.

Assumption 1. *Cross-group externalities (network effects) are not too strong, or equivalently, the platform is a live if $\alpha_c\alpha_s - 1 < 0$.*

The condition in Assumption 1 comes from the first order condition of the platform equilibrium, see A.2,which implies that the platform has positive payoff from the different transactions between providers and consumers. Assumption 1 states that the platform is sufficient over some ranges of externalities, specifically the ranges of weak externalities. As shown in the proof of the assumption,

the strong externalities, $\alpha_c\alpha_s - 1 > 0$, leads to negative profits for the platform. Thus, the platform will not enter the market if the market sides have strong externalities between each other. The reason is that strong externalities impose high performance on the platform by charging both consumers and providers with negative fees to attract and connect them on board. As a result of Assumption 1, p_c and p_s are positives. This means that the platform subsidizes the market sides by zero fee as maximum and there is no chance to subsidize them by negative fees, (i.e) the platform never pays for the market sides to attract them.

The platform faces a loop of network effects; the consumers' side depends on the number of active providers and at the same time the providers side depends on the number of active consumers. Thus, the platform requires an adequate participation of consumers to attract the providers and simultaneously it requires an adequate participation of providers to attract the consumers. The platform compensates its loss and makes profit from the subsidizing side. In this context, the question is which side the platform has to subsidize and at which level the platform should subsidize the market sides. Proposition (2) answers this question.

Proposition 2. *Under assumption 1, $B_s \leq 1$, and $B_c \leq 1$, the following hold:*

i. The platform charges consumers with positive fee below the transaction cost if $\alpha_c\alpha_s - 1 \in (-B_s(\alpha_c+1), 0)$ and with positive fee above the transaction cost if $\alpha_c\alpha_s - 1 \in (-(B_c(\alpha_s + 1) + B_s(\alpha_c + 1)), -B_s(\alpha_c + 1))$.

ii. The platform charges providers with positive fee below the transaction cost if $\alpha_c\alpha_s - 1 \in (-B_c(\alpha_s+1), 0)$ and with positive fee above the transaction cost if $\alpha_c\alpha_s - 1 \in (-(B_c(\alpha_s + 1) + B_s(\alpha_c + 1)), -B_c(\alpha_s + 1))$.

Proof. See A.3 □

Lemma 1. *Based on Proposition 2, the platform's payoff is as follows:*

i. The platform receives positive payoff if the strength of the externalities $\alpha_c\alpha_s - 1 \in (-(B_c(\alpha_s + 1) + B_s(\alpha_c + 1)), 0)$

ii. The platform always receives negative payoff if the strength of the externalities $\alpha_c\alpha_s - 1 \leq -(B_c(\alpha_s + 1) + B_s(\alpha_c + 1))$

Proof. See A.4 □

Lemma 2. *The subsidizing depends on the elasticity of demands with respect to the transaction fees charged by the platform. The platform subsidizes the providers side if the following condition is satisfied; otherwise, it subsides the consumers side.*

$$\eta^{c}_{p_c} - \eta^{c}_{p_s} < \eta^{p}_{p_s} - \eta^{p}_{p_c}$$

Proof. See A.5 □

Proposition 2 shows the size of subsidizing the platform can offer for market sides over different ranges of weak externalities. It is easy to see that the best case for consumers and providers is when their externalities $\alpha_a\alpha_s - 1 \in [max\{-B_c(\alpha_s + 1), -B_s(\alpha_c + 1)\}, 0]$ where both of them pay transaction fee below the transaction cost. However, as concluded in Lemma 1, the providers and consumers can be charged at the same time with a fee below the cost and the platform still receives positive payoff. The platform is better in the range of externalities $\alpha_a\alpha_s - 1 \in [-(B_c(\alpha_s + 1) + B_s(\alpha_c+1)), min\{-B_c(\alpha_s+1), -B_s(\alpha_c+1)\}]$ where it can charge both sides above the transaction cost. Based on this, the two-sided platform is an efficient solution if the externalities between the data providers and data consumers are weak. Lemma 1, particularly the last part of it, states that externalities must be sufficiently weak for assuring platform survivability and determine exactly the threshold that leads the platform to collapse; when the platform receives negative payoff. Lemma 1 shows that Assumption 1, which is derived from the first order condition for Equation 11 of the platform payoff, is not enough to decide about the sufficient area for the platform. Assumption 1 shows that the platform cannot sustain in the market with strong externalities between its parties while Lemma 1 proves that the platform does not sufficiently work for specific ranges of weak externalities as well.

Lemma 2 determines strictly which market side the platform should subsidize. The condition states that the providers are subsidized if the difference between their quasi-elasticities with respect to p_s and p_c are respectively greater than the difference between the consumer's elasticities with respect to p_c and p_s respectively. The consumer's elasticity with respect to p_s and the provider's elasticity with respect to p_c refers implicitly to the externalities between the consumer's and the provider's sides. The consumer is indirectly affected by the impact of p_s on the providers' demand. When p_s decreases, the number of active providers on the platform increases, which incites more consumers to exchange their transactions over the platform. Lemma 2 implies that the platform

subsidizes the providers if: 1) the consumers' demand is more inelastic (less sensitive) to the transaction fees offered by the platform compared to the providers' demand, which gives more space to the platform to play with the consumers' side and charges these consumers with higher fees; and 2) one more active provider adds more values to the consumers' demand compared to what is added to the provider's demand by one more active consumer on the platform.

2.5 Competition Case: Effects of the Direct Sale on the Platform Equilibrium

The Coase theorem does not apply to the relation between consumers and providers at the beginning of the interactions over the platform because they do not already know each other. After many transactions, data consumers and data providers share symmetric information (i.e. perfect information) about each other, where the retrial market price of data is known in the market and transaction fees imposed by the platform are known. Thus, the Coase theorem has the chance to succeed in our platform. In this section, we examine the possibility of the success of the Coase theorem applied to the relation between data providers and data consumers at long-run time (i.e. after many transactions). This case is known as unregulated because the interactions over the platform are not regulated by law.

2.5.1 Unregulated Case Description:

One main reason for users (i.e. consumers and providers) to connect to the platform is that they do not have sufficient network of connections and consequently it will be hard for them to collect the required data in a timely manner. Thus, before interacting over the platform, the consumers do not have direct access to an adequate number of providers. Similarly, before interacting over the platform, providers cannot have direct access to a large number of consumers to receive sufficient revenues from selling their data. Seeking to build sufficient networks of connections (i.e. seeking a sufficient number of consumers by providers and seeking a sufficient number of providers by consumers) is highly time-consuming and the huge costs outweigh the budgets of the consumers and providers. Thus, the consumers and providers prefer to use a platform that offers a large network of users at lower costs. The platform has different kinds of costs, including the cost of building

a sufficient network of connections. These costs are not included in the equation of the platform payoff (Equation 11) because they are only compensated at a long run-time. In other words, the transaction cost f does not include fixed costs such as the cost of building connections network.

Following a large number of transactions between the consumers and providers over the platform, they will know each other and both will have sufficient networks of connections. In other words, the main incentive for the consumers and providers to connect to the platform fades away each time they interact over the platform. Once the consumers and providers have sufficient networks of connections, they will negotiate the data trade directly without the platform, to get rid of the transaction fees imposed by the platform.

Providers are multi-home; they can directly sell their data to the consumers obtained from the previous transactions and at the same time they connect to the platform to sell data. Due to the providers' high demand, the platform has no means to control these providers and prevent them from direct sale. The platform would oblige providers not to sell the data directly, irrespective of whether the providers and consumers know each other via the platform or not. Thousands or millions of providers may use the platform to trade the data, and thus it is not sufficient or realistic to monitor them to know whether they are in compliance with their commitment (i.e. no direct trade) or not. An agreement between the providers and the platform to prevent the users from direct trade does not sufficiently work. Moreover, monitoring the users jeopardize privacy, adding costs to the platform expenses.

Consumers are single-home. However, platforms can try whenever possible to assign the same providers to the consumers. This slows down the increasing rate of the consumers' connection network. For example, let us assume a consumer i requires at least 100 providers to accomplish a sensing task. 80 providers, known by the consumer i from the previous transactions over the platform, are available to interact directly while 20 are missing. Those 80 providers are multi-home and therefore active on the platform. The consumer i connects to the platform to get 20 providers to accomplish his sensing task, but the platform might assign him 20 providers from the 80 providers. The consumer i will not benefit from this connection and he will be forced to perform 100 transactions over the platform. However, this is not the final solution to solve the problem of the direct sale (or to avoid the success of the Coase theorem). Thus, consumers are single-home,

unable to connect both the platform and the direct sale at the same time.

2.5.2 Formalization of the Unregulated Case

Data providers on our platform are often individuals supported by smart phones, raising the availability issue with respect to direct sale. For example, mobile phone users keep changing their locations during a day and they are not always available to involve sensing tasks. The platform recovers from this problem through its large connection networks that the direct sale does not have. In addition, the consumers may request different data each time, i.e. their requests change, and their connections using the direct sale may not meet their new requirements. These factors add to the expediency of the platform and prevent the success of the Coase theorem. Thus, we introduce (1) the random variable $x \in [0, 1]$ that represents the probability of each provider to perform the process of selling/buying data and meet the requirement of a new consumer request and (2) the random variable λ that represents the rate of data request performed by each consumer. Note the difference between "data request" and "transactions performed between consumers and providers": Transactions are the size of data (or number of providers) required by consumers while requests are the number of times the consumer connects to the platform asking for data. For example, the consumer i is a mobile phone sensing application that requires 3 sensing activities per day. Each sensing task requires to collect a data from 100 mobile phone users. The consumer i connects to the platform 3 times in a day ($\lambda = 3/day$), and for each connection, 100 providers (100 transactions) share their data with him.

Assume that the consumer i performs many transactions and accordingly has connections network containing $N^{s'}$ of providers. The consumer i can receive the maximum amount of data $n' = x\gamma N^{s'}$ from the direct sale. The platform has a connection network including $N_{s|q}$ providers and the consumer i can receive $x\gamma N_{s|q}$ as maximum of amount of data from the platform. However, as is explained earlier in Equation 2, the consumer's utilities do not keep increasing over the data amount. Receiving an extra amount of data, following the maximum amount of sufficient data (max_n), does not add to the consumer's utility. Thus, consumers take an amount of data $n = y\gamma x N_{s|q}$ where $y \in [0, 1]$ represents the percentage of the required data from the maximum data $(x\gamma N_{s|q})$ received from the platform.

The providers and consumers interact directly without involving the platform if the direct sale is better for both of them than over the platform. The consumers and providers pass through the platform if and only if both of them agree to interact via the direct sale. The consumers and providers incur transaction cost $f_c \geq 0$ and $f_s \geq 0$ respectively if both deiced to choose the direct sale. Note that the total transaction costs incurred by consumers and providers equals to the transactions incurred by the platform. i.e. $f = f_c + f_s$. After many transactions between the market sides (i.e. the consumers and providers), one of the sides negotiates with the other to trade directly by providing a per-transaction incentive $p_i \geq 0$. The maximum value of p_i depends on transaction fees imposed by the platform. For example, let us assume that the consumer i pays per-transaction incentive p_i to the providers to convince them to trade directly, not involving the platform. Both the consumer i and the providers agree to interact directly if and only if:

$$v_i(n|q) - n(p + p_c) \leq v_i(n'|q) - n'(p + p_i + f_c) \text{ for the consumer } i$$

$$\gamma N_{c|q}(p - p_s) \leq \gamma N_{c|q}(p + p_i - f_s) \text{ for each provider } j$$

This forces the platform to choose fees depending on p_i, not on externalities between them as in the monopolistic model. In the monopolistic case, consumers would join the platform if their utilities from using it are greater than zero while, in the presence of the direct sale, they would join the platform if their utilities from using it are greater than their utilities form using the direct sale. In other words, the cumulative distribution function that determines the number of consumers using the platform at different transaction fees in presence of the direct sale is given as follows.

$$N_{c|q} = Pr(v_i(n|q) - n(p + p_c) \geq v_i(n'|q) - n'(p + p_i + f_c)) \tag{15}$$

Similarly, the cumulative distribution function that determines the number of providers using the platform at different transaction fees is given as follows.

$$N_{s|q} = Pr(\gamma(p - p_s)N_{c|q} \geq \gamma(p + p_i - f_s)N_{c|q}) \tag{16}$$

This case is formalized as a trading selection game, which is defined as follows:

Definition 2. *Let B be the two-sided market platform, C the set of active single-home consumers and S the set of active multi-home providers. A trading selection game G is a tuple:*

$G = \langle Z, P_z, U_z(.)\rangle$.
where:

- $Z = C \cup S \cup B$ *denotes the set of players*
- P_z *is the strategies set available for each player* $z \in Z$*, defined as follows:*
  - *i.* $P_b = \{p_c^*, p_s^*\}$ *denotes the strategies set available for the platform, where* p_c^* *and where* p_s^* *are the optimal fees offered by the platform in presence of the direct sale.*
  - *ii.* $p_c = \{p_c^*, \pm p_i\}$ *denotes the strategies set available for each consumer* $i \in C$*, where* p_c^* *denotes that the consumer* c *chooses to interact over the platform and pay transaction fee* p_c^**.* p_i *denotes that the consumer* i *chooses to interact directly with providers without the platform and pays/receives incentive* $\pm p_i$*.*
  - *iii.* $p_s = \{p_s^*, \pm p_i\}$ *denotes the strategies set available for each provider* $j \in S$*, where* p_s^* *that the provider* s *chooses to interact over the platform and pay transaction fee* p_s^**.* p_i *denotes that the provider* j *chooses to interact directly with providers without the platform and pays/receives incentive* $\pm p_i$*.*
- $U_z(.)$ *represents the utility function of the player* $z \in Z$*. The utilities of players are follows:*
  - *i.* $UC_i(.)$ *represents the utility function of the consumer* $i \in C$*, where*
    $UC_i(.) = \begin{cases} v_i(n|q) - n(p + p_c^*) & \text{if } p_c = p_c^* \\ v_i(n'|q) - n'(p \pm p_i - f_c) & \text{if } p_c = p_i \end{cases}$
  - *ii.* $US_j(.)$ *represents the utility function of the provider* $j \in S$ *from selling the data for consumer* i*, where*
    $US_j(.) = \begin{cases} \gamma(p - p_s) & \text{if } p_s = p_s^* \\ \gamma(p \pm p_i - f_s) & \text{if } p_s = p_i \end{cases}$
  - *iii.* π^* *represents the utility function of the platform, where*
    $\pi^*(n) = (p_c^* + p_s^* - f)n$*.*

2.5.3 Equilibrium Analysis of the Unregulated Case

Theorem 1. *The Nash equilibrium of the game* G *is given as follows:*

(1) The best response of the platform is given by the following imposed fees:

$$p_s^* = f_s - p_i \tag{17}$$

$$p_c^* = \frac{v_i(n|q) - v_i(n'|q) - np + n'(p + p_i + f_c)}{n} \tag{18}$$

(2) The best response of the consumer i given p_c^* in Equation 18 is paying p_c^* and trading via the platform.

(3) The best response of providers given p_s^* in Equation 17 is paying p_s^* and trading via the platform.

Proof. See A.6 □

Corollary 1. *The platform is able to attract consumers and providers on board if and only if the amount of data n provided by the platform is strictly greater and more sufficient than the amount of data n' provided by direct interaction.*

Proof. See A.7 □

Theorem 1 characterizes the Nash equilibrium of game G that represents the unregulated case. Corollary 1 implicitly states that the Coase theorem will not apply to the relation between the data consumers and data providers as long as the platform is able to provide more sufficient data for the consumers compared to the direct interaction. Thus, the Nash of G implies that the platform provides a larger amount of data, but providing larger amount of data does not imply Nash (interacting via the platform). Providers would always sell their data directly when they receive more utilities by getting rid of transaction fees imposed by the platform. When consumers reach a sufficient size of network of connections, consumers and providers can negotiate the fees of the platform to trade the data directly without the platform. The following question is raised in this context: "To which extent the platform will resist the Coase theorem and stay able to provide more sufficient data?" Lemma 3 answers this question.

Lemma 3. *The relation between the consumer i and the interacting providers will satisfy the Coase Theorem when the consumer i performs r fulfilled requests of buying data. The number r is given as follows:*

$$r = \frac{2(1-x) - \frac{max_n + \Delta}{max_n}}{\log(1-x)} \tag{19}$$

where:

- $x \in [0, 1]$*: a variable that represents the stability of providers in the platform.*
- max_n*: the maximum sufficient amount of data that the consumer i requires.*

- Δ*: A small tolerance value that represents the difference between the actual amount of data that the consumer* i *receives when he makes* r *requests and the maximum sufficient amount of data that the consumer* i *requires.*

Proof. See A.8. □

The success of the Coase theorem depends on the stability of providers and the sufficient amount of data required by the consumers. As stated in Lemma 3, the Coase theorem will succeed after consumers perform a certain number of requests for buying data. Table A.1 shows values of r over different ranges of x and (max_n) combined with a certain number of requests that consumers performed per day. As noted in the table, the success of the Coase theorem engrosses long time when providers have a low stability in the platform (small values of x). By contrast, the success of the Coase theorem is inevitable when providers have a high stability on the platform (high values of x). Based on these observations, we can conclude that monetizing data process by the two-sided market requires a regulation in the domains where the providers have a high stability. The regulation process is related to a set of procedures that prevent the success of the Coase theorem and protect the platform from collapsing. In this context, the following two questions are raised: "In which domains we need a regulation? What are potential procedures for the platform in the regulated case?" Section 2.7.2 answers both questions.

2.6 Simulations and Empirical Analysis

In this section, we evaluate the two-sided market model over an alternative model for data monetization: the classical form of intermediaries (i.e., merchant model). Specifically, we compare the total surplus of the involved parties under the two-sided market with the total surplus under the merchant model. Then, we validate the efficiency of the two-sided market model over different ranges of the users stability. Figure 2.4 shows the simulation inputs, adjustable parameters, and outputs. The simulation aims to achieve the following objectives:

- **Objective 1:** Checking the efficiency of the two-sided market against the merchant model. Specifically, we simulate the total surplus (i.e., consumers payoff, providers payoff and platform payoff) under both the two-sided market model and the merchant model. The simulation

increases gradually the externalities (i.e., the parameters α_c and α_s) and then calculates the total surpluses of the involved parties. The simulation aims to check how much the two-sided market can contribute in terms of payoff of the involved players over different ranges of externalities.

- **Objective 2:** Checking the impact of the users stability on the platform payoff. The simulations increases gradually the users stability (i.e., the parameter x) and then calculates the platform payoff.

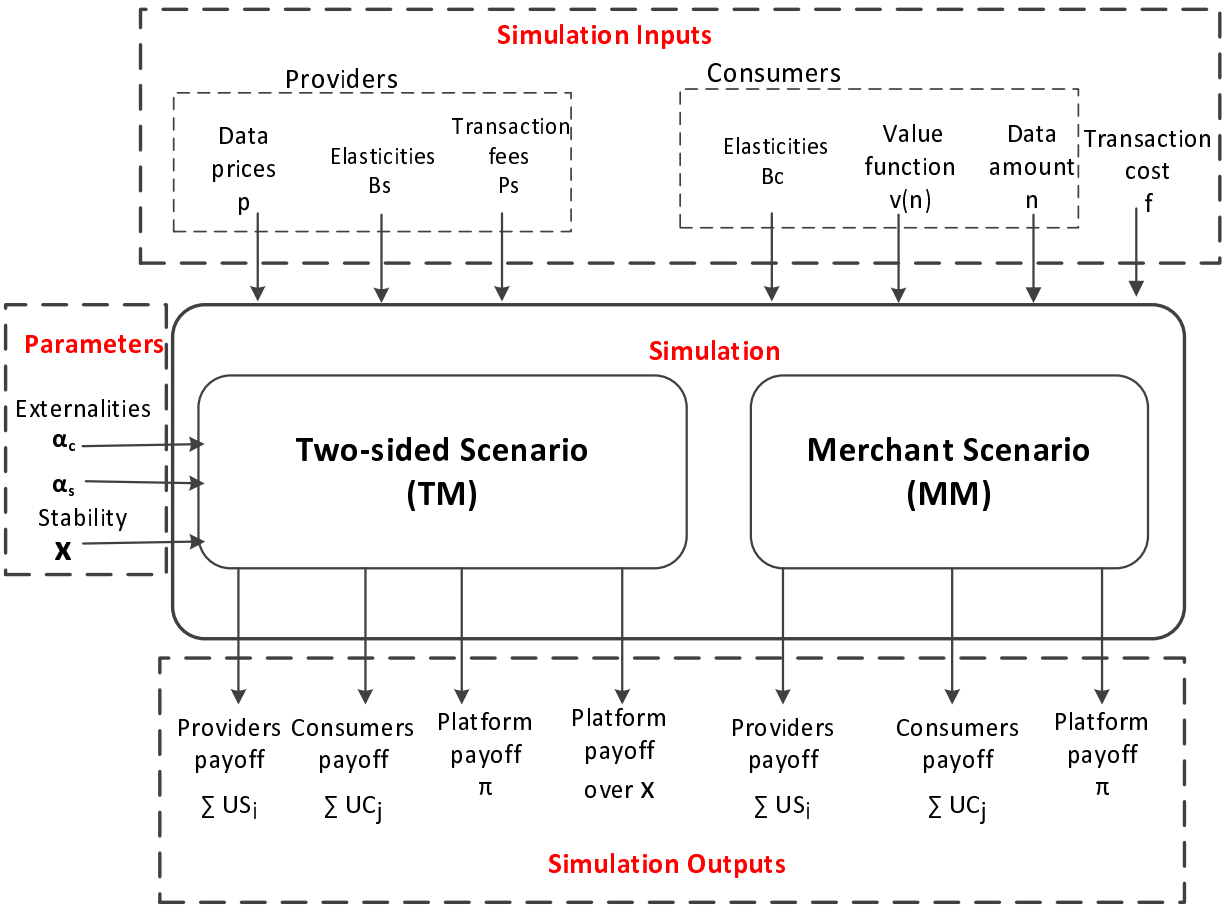


Figure 2.4: Simulation overview

The simulation parts, as shown in Figure 2.4, are described in the following subsections: 2.6.1, 2.6.2, 2.6.3, 2.6.4 and 2.6.5.

2.6.1 Simulation Inputs and Parameters

In this section, we describe inputs and parameters of the simulation shown in Figure 2.4. Specifically, we describe the used data and associated demand distributions. Unfortunately, there is no available dataset about personnel data in the context of monetizing data. However, we use a closest real dataset provided in [26]. The dataset contains $1,025,908$ market transactions involving $1,641$ of the most popular products on amazon. Product prices in the dataset are close to each other, and are normally distributed with a mean of \$21 and standard deviation of \$17. Each market transaction in the dataset contains the retrial price, seller ID, seller rating, transaction date, product type, shipping price, and the lowest price in the market at the transaction time. Although the dataset is about the market transactions of commercial goods instead of commercial data, we are still benefiting from prices' real distributions under real supplying levels.

A hundred samples of 10000 market transactions at different times are used as inputs. To guarantee an accurate simulation, we used a confidence level of 95% with regard to the price mean for taken samples. Sellers act as data providers in the simulation and product prices represent their utilities while shipping prices represent transaction costs (f). Each taken sample contains data providers $(N_{s|q})$ that are distributed normally with mean 1000 and standard deviation 100. According to [26], amazon charges sellers in average 15% of product price (i.e., P_s). The providers and consumers elasticities are set up between $0.1 - 0.3$ (i.e., β_s and β_c), which are similar to the sensitivity of mobile/telecommunication services price shown in the literature [31]. The intuitive value for the externality from the provider's perspective (α_s) varies from 0.1 to 0.9 since one consumer contributes by less than one to the provider demand. The same intuition applies to the externality from the consumer's perspective α_c. The consumer demand, $N_{c|q}$, is then computed in Equation 6. The amount of data n required by each consumer is distributed normally with a mean of 500 providers and a standard deviation of 100 providers, which matches the available number of providers in the taken samples. We assume that the value function obtained by consumers from each transaction is normally distributed with mean 42 and standard deviation 17. The value function distribution is similar to the data price distribution with a larger mean, which guarantees reasonable ranges of consumer utilities. Table 2.2 shows the values of the system parameters.

Simulation Parameters	Values
$N_{s\|q}$	$N \sim (1000, 100)$ data providers
P_i	$N \sim (21, 17)$ USD
f	50 - 100 cents
α_s	0.1 - 0.9
α_c	0.1 - 0.9
γ	1
β_s	0.1 - 0.3
β_c	0.1 - 0.3
p_s	15%
n	$N \sim (500, 100)$ data transactions
$v(n)$	$N \sim (42, 17)$ per transaction

Table 2.2: Simulation Parameters Values

2.6.2 Two-Sided Market Scenario

In this section, we describe the two-sided scenario shown in Figure 2.4. The two-sided market platform (TM) is given by the following scenario:

(1) Providers reveal their data prices and their availabilities at different times.

(2) Consumers reveal their interest at different data price levels including the required data amount and their utilities.

(3) The platform observes the consumers and providers utilities.

(4) The platform calculates the optimal charging fee (p_c) that maximizes its payoff as given by Equations 11.

(5) The platform reveals the charging fees to consumers and providers.

(6) Once the charging fees are accepted by consumers and providers, they start interacting with each other directly and setup data trading.

2.6.3 Classical Intermediary Scenario

In this section, we describe the merchant scenario shown in Figure 2.4. The main difference between the classic form of market intermediaries (merchants) and the two-sided models is that

pure merchants take possession of the sellers' goods (i.e. buying sellers' products) and resell them to consumers at retrial prices. By contrast, the two-sided platform leaves entirely the selling process to the sellers and buyers and simply determines the buyer's and seller's fees with a common marketplace [11]. Under the merchant mode, the intermediary buys the provider's data by offering a buyout bid B^s for each seller and resells the data to the consumers for a retrial price p'. Each data provider only cares about the bid he is biding with, not about the number of active consumers connected to the other side of the platform. Assume $p(n|q)$ as the price function of the data amount n given the quality q, which the merchant will pay for the providers after accepting their bids B^s. Assume that there are $N_{c|q}$ consumers in the market and each consumer i has utility u_i and requires n_i amount of data. The merchant then receives payoff as given in Equation 20. Comparing to the two-sided platform, the merchant offers the best $B^s \leq (p - p_s)N_{c|q}$ and the providers then receive a total surplus as given in Equation 21.

$$\pi_M = \sum_{i=1}^{N_{c|q}} p'(n_i|q) - p(n_i|q) \tag{20}$$

$$TS_M = p(n) = \sum_{i=1}^{n} B_i^s \tag{21}$$

Based on Equation 20, the merchant maximizes his payoff by maximizing p' and minimizing $p(n)$ as much as possible. Data differ from other economic goods for the possibility to be resold to many consumers at the same time. In the merchant model, because of selling economic goods that cannot be resold, the equilibrium of the market is given by the intersection of the demand curve and the supply curve, which means the quantities of goods needed by the consumers equal the quantities of goods provided by the sellers. This is not true in our case since the same data can be shared between all consumers. For instance, if there are 1000 consumers and each consumer requires 100 units of a particular data, it does not mean that 100000 data units are needed, where it is enough for the merchant to buy 100 units and share them with the whole data consumers. Thus, Equation 20 is updated as given by Equation 22, where $p(n|q)$ is the price of total data that will be shared among the consumers. This will largely raise the competition among data providers and push them to accept lower bids $B^s << (p - p_s)N_{c|q}$, which negatively affects their total

surpluses. The merchant can create more values and extract more profits from the data providers. Unlike the merchant, the two-sided market platform maximizes its payoff (given in Equation 11) by maximizing the total transactions among the providers and consumers which positively affect the provider's total surpluses. Furthermore, the two-sided market guarantees the surplus distribution between larger data providers, decreasing the competition between the providers and creating more values for the consumers. For instance, if there are 1000 consumers and each consumer requires 100 units of a particular data, the two sides distributes 100000 transactions over all providers, while only 100 providers (if each provider provides one data unit) will share the total provider's surplus under the merchant model. As explained earlier, a huge number of data providers are not involved in the process of monetizing data because of the unworthy and insignificant rewards. The two-sided model helps "increase" the percentage of providers by creating more surplus offered to the provider's sides.

$$\pi_M = (\sum_{i=1}^{N_{c|q}} p'(n_i|q)) - p(n|q) \tag{22}$$

The Merchant model (MM) is given by the following scenario:

(1) Consumers send data requests to the platform including the required data amount and their utilities.

(2) The platform sends data requests for providers.

(3) The providers send their availabilities and their prices.

(4) The platform calculates the optimal data prices and the optimal data amounts (maximize Equation 22) and offers them to both sides.

(5) Providers and consumers can accept or reject the platform offer based on their utilities.

(6) The platform buys the data from the providers who accept the offer.

(7) The platform sells the data to the consumers who accept the offer.

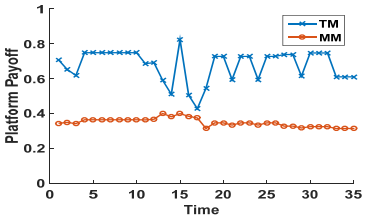
2.6.4 Simulation Results: Two-Sided Market vs Classical Intermediaries

In this section, we describe the simulation outputs. We run our simulation over different ranges of the externalities ($\alpha_c\alpha_s$) at 0.35, 0.45, 0.55, 0.65, 0.75 and 0.85. The values of externalities are given to the simulation as inputs. According to those values, the simulation adjusts the consumers and providers utilities and then executes above scenarios. The simulation results are given in terms of platform's, consumers' and providers' surpluses over the time in days as shown in the Figures 2.5, 2.6, and 2.7 respectively. For example, Figure 2.5a describes the platform payoff at externalities 0.35 over 35 days of transactions between the providers and consumers. The platform payoff is normalized in the figure with respect to the maximum payoff received by the platform during the run time of the simulation at different levels of externalities. As shown in the figure, in the fifth day, the platform receives 0.7 as a payoff under the two-sided market (TM) while it receives 0.4 as a payoff under the merchant model (MM). Figures 2.5, 2.6 and 2.7 are interpreted in the same manner.

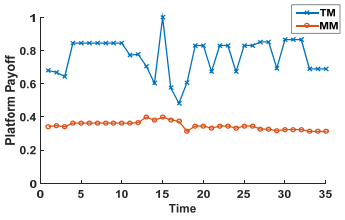
As noted in those figures, the involved parties receive in general higher payoff under the two-sided market model. Specifically, the two-sided market show more efficiency than the merchant model under week externalities (0.35, 0.45, 0.55 and 0.65) where the platform and the providers receive higher payoff as shown in Figures (2.5a - 2.5d) and Figures (2.7a - 2.7d) respectively. However, the two-sided market shows less efficient outcomes as the externalities become stronger. Specifically, providers' surpluses decrease gradually as the externalities is increased. For example, the total surpluses of providers in Figures 2.7a at externalities 0.35 are relatively higher than the total surpluses in Figures 2.5b at externalities 0.45. This decreasing is continuing until the two-sided market show almost the same efficiency of the merchant model under the strong externalities (0.75 and 0.85) as shown in Figures 2.7e and 2.7f. Similarly, the platform receives less surpluses as the externalities become stronger. For example, the platform's surpluses are relatively higher at externalities 0.45 than the surpluses at externalities 0.55 as shown in Figures 2.5b and 2.5c respectively. This decreasing is continuing until the two-sided market show the same (as shown in Figure 2.5e) or less efficiency (as shown in Figure 2.5f) than the merchant model. The reason behind this, according to Proposition 2, the platform imposes slight transaction fees (less than transaction costs) to attract both sides at the strong level of externalities, which leads to less payoff.

The consumers receive always higher payoff under the two-sided market model as shown in Figures 2.6a - 2.6f. The reason behind this is that the platform makes profit from both sides (providers and consumers) under the two-sided market model by imposing transaction fees on both sides while it makes only profit from the consumers side under the merchant model, i.e., the providers bear a part from the desired profit that the platform wants under the two-sided market. Moreover, the consumers purchase the data at lower price from the providers under two-sided market while in the merchant model, the consumers purchase the data from the platform (broker) at higher price. Unlike to the platform and providers, the consumers' surpluses increases gradually as the externalities become stronger. For example, the consumers' surpluses at externalities 0.75 shown in Figure 2.6e are relatively higher than surpluses at externalities 0.65 shown in Figure 2.6d. Similarly, surpluses at externalities 0.65 shown in Figure 2.6d are relatively higher than surpluses at externalities at 0.55 shown in Figure 2.6c. The reason behind this, according to Proposition 2, the platform impose less transaction fees as the externailities become stronger to attract the consumer side, which leads to less cost bearing by the consumers. However, the consumers' total surpluses become less under too strong externalities (0.85) shown in Figure 2.6f. The reason behind this is the high impact of the strong externalities on the subsidy technique that leads to less attraction from consumers and providers to sustain a positive platform payoff.

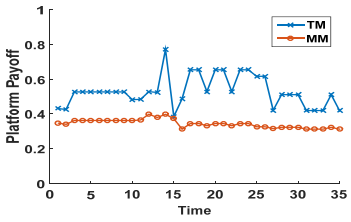
Also noted in those figures, the payoff of the two-sided market fluctuates largely than the payoff of the merchant model. For example, in Figure 2.5a, the changes of the platform payoff over the time (10 - 20) are clearly noted in the two-sided market curve. The reason is that the structure of the two-sided payoff rely on all instant individual transactions performed among the users, which leads to a full market coverage on the consumer side, while the structure of the merchant payoff rely on the competitive price that maximizes the total revenues transferred to the platform, which leads to partial market coverage on the consumer side. Thus, the total changes are larger in the case of the full market coverage in case any changes happen on products prices, consumers and providers utilities, ..etc. Therefore, the two-sided market is more responsive to the changes such as product price changes than the merchant model.



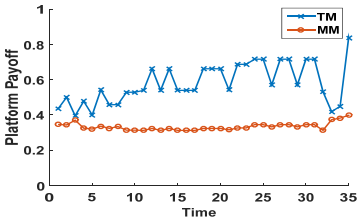
(a) Platform payoff at $\alpha_c\alpha_s = 0.35$



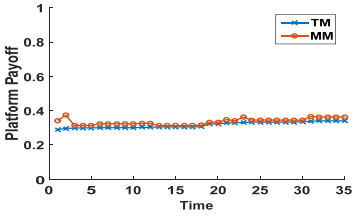
(b) Platform payoff at $\alpha_c\alpha_s = 0.45$



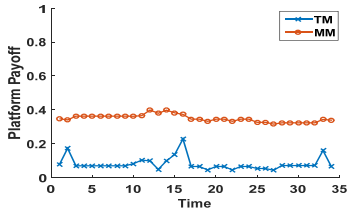
(c) Platform payoff at $\alpha_c\alpha_s = 0.55$



(d) Platform payoff at $\alpha_c\alpha_s = 0.65$

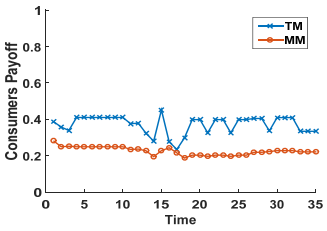


(e) Platform payoff at $\alpha_c\alpha_s = 0.75$

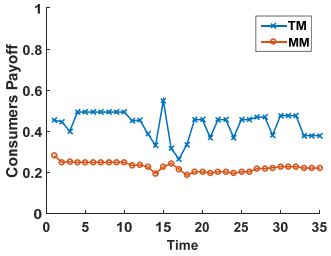


(f) Platform payoff at $\alpha_c\alpha_s = 0.85$

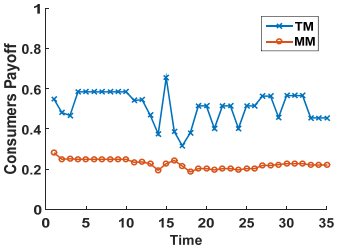
Figure 2.5: Platform payoff over two-sided and merchant model



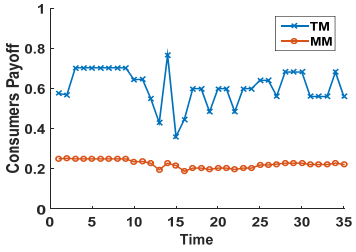
(a) Consumers payoff at $\alpha_c\alpha_s = 0.35$



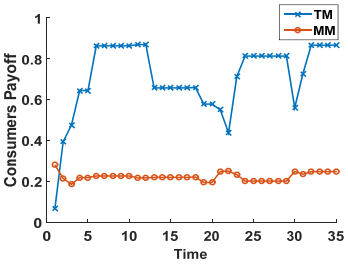
(b) Consumers payoff at $\alpha_c\alpha_s = 0.45$



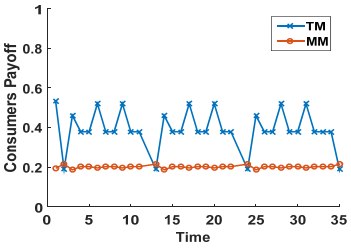
(c) Consumers payoff at $\alpha_c\alpha_s = 0.55$



(d) Consumers payoff at $\alpha_c\alpha_s = 0.65$

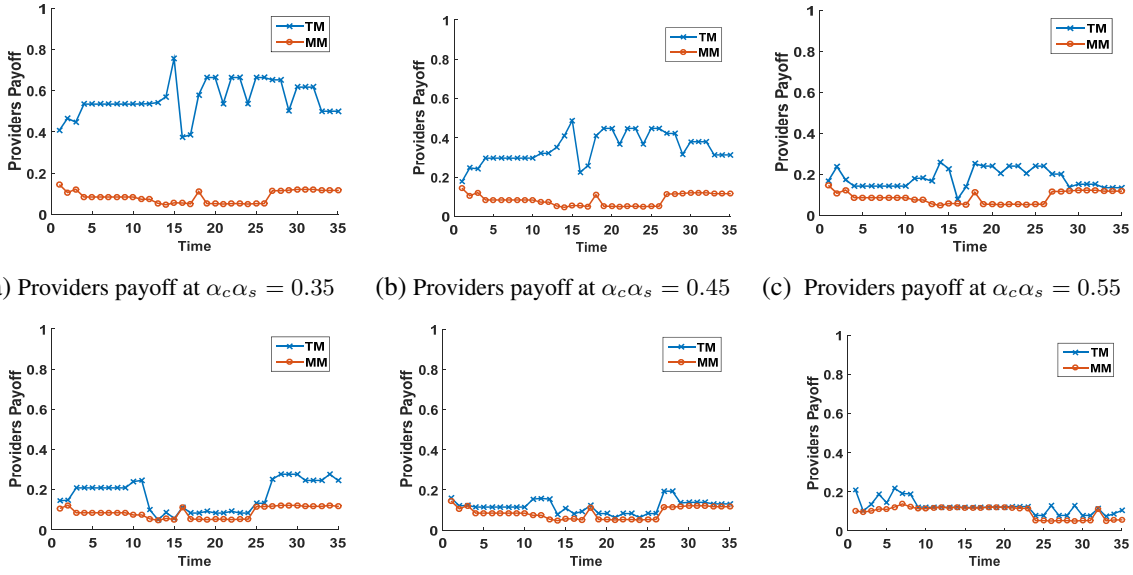


(e) Consumers payoff at $\alpha_c\alpha_s = 0.75$



(f) Consumers payoff at $\alpha_c\alpha_s = 0.85$

Figure 2.6: Consumers payoff over two-sided and merchant models



(a) Providers payoff at $\alpha_c\alpha_s = 0.35$ (b) Providers payoff at $\alpha_c\alpha_s = 0.45$ (c) Providers payoff at $\alpha_c\alpha_s = 0.55$

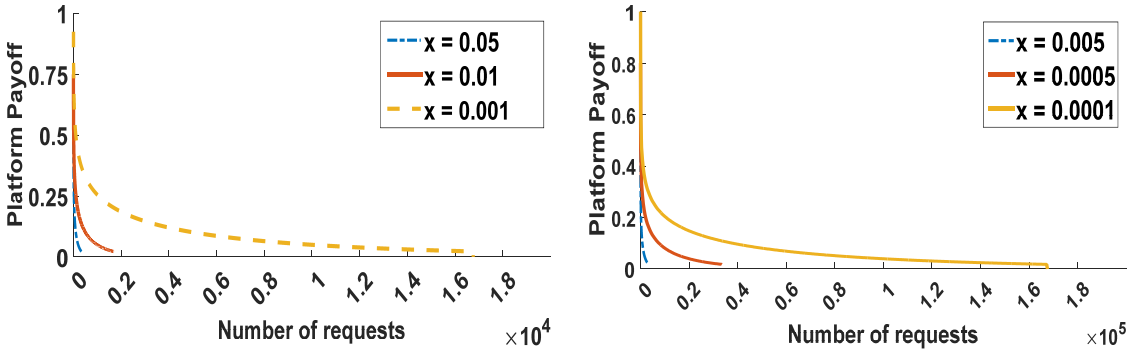
(d) Providers payoff at $\alpha_c\alpha_s = 0.65$ (e) Providers payoff at $\alpha_c\alpha_s = 0.75$ (f) Providers payoff at $\alpha_c\alpha_s = 0.85$

Figure 2.7: Providers payoff over two-sided and merchant models

2.6.5 Simulation Results: Two-Sided Market Efficiency Over Users Stability

In this section, we continue describing the simulation outputs. Specifically, we validate the efficiency of the two-sided platform over different ranges of providers availability. We extend the simulation application provided in the previous section by including a new parameter representing the provider's availability. The parameter is normally distributed and takes values between 0-1. Once a provider and a consumer interact with each other over the platform, the simulation records that the provider and the consumer know each other, and they become able able to pass the platform. The platform adjusts its strategy according to this scenario and calculates the optimal fees. The simulation results, shown in Figures 2.8a and 2.8b, are given in terms of the platform payoff and the number of data requests performed by each consumer. As noted in the figures, the two-sided market shows efficient outcomes under low levels of the users availability and less efficient outcomes under higher availability levels. In other words, the platform payoff falls down relatively faster under high levels of the users availability. In Figure 2.8a, at the high level of providers stability ($x = 0.01$), the platform receives zero payoff after consumers perform 2000 data requests, i.e., consumers and providers pass by the platform after 2000 data requests from the consumers side. In Figure 2.8b, at the low level of providers stability ($x = 0.0001$), consumers and providers pass by the platform

after executing 160000 data requests from the consumers side.



(a) Platform payoff at high levels of providers stability (b) Platform payoff at low levels of providers stability

Figure 2.8: Platform payoff at different levels of providers stability

2.7 Discussion

Recently, some data monetizing platforms have been launched by industrial communities such as People.io[5], Opiria[6], and Lotame[7]. Those platforms provide a secure digital marketplace where people can monetize and trade their personal data. These platforms aim to help individuals control with whom they share specific personal data and get proper compensation. In parallel, they help data consumers such as publishers, marketers and agencies find new data providers, increase engagement, grow revenue, and get easy and quick access to personal data of high quality. Those platforms provide unstacked data solutions for large domains such as marketing, health care and technology. They involve different types of data and serve different enterprises worldwide. In this section, we discuss 1) how the type of data affects the behavior of such platforms; 2) in which domains such platforms exhibit tangible efficiency; and 3) what are the potential limitations of such platforms from the business perspective.

2.7.1 Effect of Data Nature on the Subsidy Technique

In Section 2.4.6, specifically Lemma 2 shows that subsidizing the market sides relies on the elasticities of their demands. As known in economics, the demand's elasticities are controlled and

[5] https://econsultancy.com/start-me-up-people-io-allows-people-to-monetize-their-personal-data/
[6] https://medium.com/@EVALUAPE1/blockchain-project-review-opiria-6-3-dapp-for-data-exchange-fdc78a8be7b7
[7] https://www.lotame.com/its-time-to-unstack/

affected by many factor-related the characteristics of the users (i.e. the buyers and sellers), the nature of goods and business and the environments in which the users interact. In this section, we review those factors and connect them to the process of monetizing data. Specifically, we discuss how the nature of data affects subsidizing the consumers and providers.

(1) Characteristics of providers and consumers:

The characteristics of the users are summarized by their budgets and the size of their business. The low budgets of the consumers restrict their purchasing power and make them unable to pay for the data even if these data contribute largely to their utilities. This increases the elasticity of the consumer demand with respect to the charged fee and makes the consumers more responsive to the fees changes. By contrast, consumers who have large budgets, do large businesses and relatively make large profits are less responsive to the fees' changes. When the platform increases transaction fees, rich consumers (with large budgets and large businesses) will still be willing to connect to the platform and perform the same amount of the transactions, while poor consumers (with low budgets and small businesses) will not be able to perform the same amount of transactions over the platform. Similarly, the elasticity of the providers' demands is restricted by the size of their budgets and businesses. When the providers are individuals, often with very low budgets, low price for data and small businesses relatively achieving nominal profits, their demand curve shows highly elastic behaviors. By contrast, the demand curve of the providers shows inelastic behavior when the providers are professional organizations specializing in the process of data trade.

As a result and informally, when individuals attend the providers side and professional organizations attend the data consumers' side, the platform tends to subsidizes the data providers side and make profit from the consumers' side. The reason behind this is that consumers are relatively more able to bear larger transaction fees without affecting the number of performed transactions, while any changes in the transaction fee offered to the individual providers will relatively lead to big changes on the number of performed transactions. Similarly, the platform subsidizes the data consumers' side and makes profit of the providers side when the providers reach the level of professional organizations and the consumers have low budgets

and small businesses.

(2) Rare vs abundant data

Rare available resources of data induces the consumers to show less resistance against increasing transaction fees because they cannot switch to other resources. Thus, the demand of data consumers shows inelastic behavior in the case of rare data. This gives the platform the chance to increase the transaction fee offered to the consumers without affecting the number of performed transactions. On the other hand, the demand of consumers depends on the size of the provided sufficient data. Thus, the platform is forced to highly subsidize the providers side with very nominal or zero transaction fee to guarantee the maximum level of sufficient data. By contrast, the consumers' demands show a highly elastic behavior with respect to the transaction fees imposed by the platform in the case of abundant data; the consumers can easily switch to other data resources even if there are only small changes in the transaction fee because several substitutes of data resources are available. Thus, the platform shows more flexibility and imposes less transaction fees on the consumers' side rather than in the case of rare data. On the other hand, the data abundance leads to aggressive competitions between the providers and induces them to accept higher transaction fees imposed by the platform.

(3) Real time vs historical data

Many data consumers provide real time services such as traffic map applications [65, 41, 94, 40, 45, 7]. Those services require real time data, making the data providers an urgent need for the consumers and restrict their abilities to postpone the consumption of data. The platform uses the urgency of the consumers and maximizes its profit by imposing higher transaction fees without affecting the number of performed transactions. Consumers find it difficult to shift to other substitutes in a short period, in order to respond to a change in transaction fees. Thus, the consumer's demand shows inelastic behavior in the case of the real time data. On the other hand, the data providers use the urgency of consumers and put more pressures on the platform by postponing the sales process until the expiration as in the case of stockholding strike. This forces the platform to show more flexible behavior and highly subsidizes the providers' side to guarantee sufficient data on time. However, the time

factor may also negatively affect the elasticity of the providers' demand as follows: After a certain amount of time, no consumer becomes interested in this kind of data and then its value drops to zero. Thus, data providers adjust their behaviors over the data expiry and discount perishable data as their expiry date approaches, in an attempt to reduce waste. This kind of behavior applies to cases with abundant providers and they just want to disposal from their data in the nearest possible time, which leads to inelastic providers' demands. This case gives the platform the chance to impose higher transaction fees on the providers' side. By contrast, consumers who require less updated data (historical data) have highly elastic demand as their consumption can be postponed in the case of an increase in the transaction fees imposed by the platform. This forces the platform to charge consumers lower transaction fees compared to the case of real time data.

(4) Sensitive vs non sensitive data

Sensitive data includes all data that may breach user privacy when data is disseminated. Specifically, privacy concerns arise when the platform requires to reveal the providers' identities combined with confidential attributes including private information such as salaries, medical tests and sexual orientations. Many data providers are not comfortable and not willing to share their identities combined with private information. This concern may affect the availability of data. The platform can use the concerns of the providers and maximizes its payoff by increasing transaction fees imposed on the consumers since they cannot switch to other data resources. Consumers have no choice and accept the increase of transaction fees without performing less transactions. The consumer's demand shows a high inelastic behavior with respect to transaction fees, where consumers perform the same amount of transactions at different transaction fees. On the other hand, the platform has to perform efficiently to convince the maximum sufficient number of providers to reveal their data. Thus, the platform shows more flexibility with the providers and highly subsidizes them by offering zero transaction fees. By contrast, the platform charges providers with higher transaction fees and makes less profit from the consumers' sides in the case of non-sensitive data.

2.7.2 Efficiency of the Two-Sided Platform in Collecting and Sharing Data

As mentioned earlier at the beginning of this section, recently launched platforms serve different types of enterprises. Such a variety of served enterprises raises the research question "how the business type of served enterprises affects the platform?"

In Section 2.5, we discussed the effect of the direct sale on the performance of our two-sided platform. Lemma 3 and the simulation results in Section 2.6.5 show that the success of the platform depends principally on the stability of the providers in the system, where the platform achieves higher profits and continues providing its services as long as the probability of the providers meeting the consumers' requirements over many data requests is low. In this section, we discuss the providers stability in some domains of collecting and sharing data.

In the field of mobile phone sensing, the providers are mobile phone users (as movable points), which leads to an availability issue. The availability of spatio-temporal data of mobile phone users has been studied extensively by a large number of researches such as [36, 91, 30]. [19] surveys the contributions made so far on the mobile phone sensing applications and social networks that can be constructed with data sets, the study of personal mobility, geographical partitioning and urban planning. According to [36] observations, the probability of finding a mobile phone user in a specific location ranges from 1×10^{-6} to 1×10^{-2}, which falls (based on table A.1 in Appendix G) within the acceptable range required for the success of our platform.

In the domain of marketing, enterprises desperately need personal data from data providers to understand their needs in order to design products that fuel consumers' desires and perfectly meet their requirements. In such domains, the provider's location is incidental relatively, and hence the provider shows middle range of stability. According to Lemma 3, monetizing platforms may face a possibility of collapsing in such cases. The health care domain is another example where patients can also show middle range stability in the system and they can directly provide updated medical records for interested data consumers skipping the monetizing platform. However, this situation has less impact because health care applications/studies are generally interested in larger size of patients data, which can be appropriately accomplished by the interactions via the platform.

Many big organizations (the second data owners) are involved in the process of buying and selling data. They collect data about their users, their habits, their mobility and their acquaintances and they share information for market purposes and for gaining competitive advantages. These organizations are widely seen in the domain of viral marketing, crime detection and health care services. Acxiom, LexisNexis, ChoicePoint, Equifax, Experian, TransUnion and the federal agency for Medicare and Medicaid Services are famous organizations that engage in the process of collecting and sharing data. Such organizations have high stability in the system and each one is able to provide tremendous amount of data. Thus, the platform, according to Lemma 3, faces a high possibility of collapsing in the unregulated case.

It is finally concluded that the efficiency of the platform depends on the nature of the domains and the parties involved in the process of collecting and sharing data. The platform is efficient and still receives positive payoff: (1) in dynamic environments where the providers show low probability to involve many and different collecting data/sensing tasks; (2) in domains that require relatively a large number of providers; and (3) in active and renewed environments where the curve of platform users grows relatively quickly. In such environments, consumers will not benefit from the direct sale and still need to connect to the platform to use its users network. Once there is high providers stability, the platform should prevent the consumers and providers from knowing each other, for instance by interacting anonymously or via indirect sale if possible. In other words, the platform should take the data from the providers and deliver it to the consumers.

Interacting anonymously between data providers and data consumers is an applicable and practical solution. Opiria[8], for example, uses the blockchain technology to guarantee a protection for data providers' anonymity. Opiria enables individuals to create a passive income stream by monetizing their personal data over the Ethereum blockchain. Furthermore, interacting anonymously is supported by the General Data Protection Regulation (GDPR)[9] that focuses on preserving the privacy of data providers. GDPR recommends to adopt encrypted mechanisms that hide the provider's identity before sending her data to the consumers. Such encrypted mechanisms are seen in the domain of mobile phone sensing applications [92, 69, 32, 85]. This means consumers interact with

[8] https://medium.com/@EVALUAPE1/blockchain-project-review-opiria-6-3-dapp-for-data-exchange-fdc78a8be7b7
[9] https://gdpr-info.eu/

anonymous – but reliable – providers and the platform does not give the chance to the consumers and providers to know each other. It is finally concluded that preserving the privacy of data providers contributes positively to data monetization platforms. In fact, international and law regulations such as GDPR have positive impact and protect not only data providers, but also sustain the business of data monetization platforms.

Indirect interaction is another potential solution. Thus, instead of linking providers to consumers so they can interact directly, the platform identifies the right providers, but sell itself their data to the consumers. Indirect interaction (the platform selling the data) does not release the platform from the effect of two-sided-market since the providers still set their prices and still sell their data at retail prices, not at wholesale prices as the merchant model. The role of the platform in the indirect sale is limited to take the data from the providers side and deliver them to the consumers side, not to purchase the data from the providers and sell them to consumers. Thus, the platform will not collapse and resolves the problem of the unregulated case. The indirect interaction solution can be used in real-time sensing activities, as an example, forcing the platform to follow indirect sequence of interactions between the providers and consumers to guarantee data delivery on time.

2.8 Related Work

In this section, we discuss relevant work related to the two-sided model and data market. In Section 2.8.1, we review the related work with regard to the two-sided market while we discuss the differences between our work and the original model of two-sided market in Section 2.8.2. In Section 2.8.3, we review the literature about the data market.

2.8.1 Two-Sided Market Literature

Since early 2000s, there has been a great research towards two-sided market theories, which introduced the subsidy strategy to maximize intermediary payoffs by imposing a price lower than the marginal cost on the users of one side and creating values for the users of the other side. In their seminal contribution to the literature of the two-sided market, [82, 83] state that the outcome is better by setting the price structure not the total price level to get all sides on board. They focused on the transaction-based markets such as credit cards and payment systems. The second key paper

in the literature of the two-sided market [12] studies the equilibrium price for the two-sided market. The author argues that the equilibrium varies based on three factors: (1) the size of the cross-group externalities between market sides; (2) the form of the imposed fees; and (3) whether users are single-home or multi-home.

Many applications have been studied extensively based on the two-sided market theories. [34] discuss network neutrality regulation of the Internet in the context of a two-sided market model. The authors show that the total surplus has increased under network neutrality regulation compared to positive fees imposed on the content providers. [73] provide a game-theoretic model based on a two-sided market framework to examine social welfare under neutral and non-neutral network platforms. The investigators found that the non-neutral network is always welfare superior in a "walled-gardens" model while the neutral network is superior in a "priority lanes" model when CP-quality (content providers quality) heterogeneity is large. [49] discusses electronic intermediaries when trading partners are involved in a commercial relationship, based on the two-sided market literature that discuss both efficient pricing and monopoly pricing. [64] constructs a two-sided market media platform model. The paper demonstrates platform equilibrium under three factors concerning the context of sharing media: matching technology, prosumer strategy and advertising technology. [24] analyze the impact of mergers on prices imposed on both newspaper subscribers or advertisers in the Canadian newspaper industry. They found that greater mergers did not lead to higher prices for either newspaper subscribers or advertisers. [87] studies empirically the externalities in the yellow pages directories. However, like all other relevant contributions, the users side of the directories does not pay (i.e. the price is zero). However, most of the published papers in this context have assumed either linear demands or zero price on one of the market sides. Unlike those proposals, we consider in our analysis a non-linear demand and we examine the platform equilibrium over non-zero prices. Moreover, we are the first who provide an analysis for the monetizing personal data using the two-sided market model.

2.8.2 Comparison with Rochet's [82] and Armstrong's Work [12]

We grounded our analysis on Rochet [82] and Armstrong [12]. However, there are a number of modeling differences between our article and their analysis. These differences are as follows:

i. Non-linear externalities: [82] and [12] proposals use linear externalities, which do not reflect the realistic case, while we use logarithmic functions. Specifically, we differ from them by Equations 3, 4, 5 and 6. In the linear model combined with per-transaction fees, as proposed by [82], the users' incentives to join the platform do not depend on the platform's performance on the other side and these users will join if and only if the utility received from a transaction is greater than the fees imposed by the platform. This cannot be applicable to our platform. The main incentive for consumers to join the platform is getting an adequate amount of data. Successful transaction (i.e. the utility of one data unit being greater than the transaction fee) is necessary but not enough for the consumers to join the platform. The size of successful transactions (i.e. the data amount) is the main condition for a consumer to join the platform. For example, consumer i needs at least 100 providers to accomplish a certain sensing task. The Consumer is not willing to join the platform if the platform matches less than 100 provider for his request. The consumer i first checks whether the platform is able to provide 100 providers or not. [12] uses the linear externalities combined with fixed affiliation fees. In his proposal, the user's incentives to join the platform depend explicitly on the externalities between the market sides. However, he assumes that a user benefits from all users on the other side, and thus, it is not applicable to our platform since a consumer does not buy the whole data provided by all the providers. We deal with these issues in our proposal by the logarithmic model which reflects the realistic behavior of the data users. Our modification has major effects on the price structure for the two-sided market. Specifically, Proposition 1, Equations 9 and 1 for profit maximizing are different in our paper. Moreover, we extract the subsidy condition 2 that determines exactly which side the platform has to subsidize.

ii. Direct sale: Both articles propose assumptions to make the Coase theorem not applicable to the relation between market sides and they role out the direct interaction between market sides by imposing limitations on the interactions between market sides. Those limitations imply that 1) the providers cannot offer two different prices for data depending on whether the consumer purchases directly from the provider (direct sell) or by the platform; in other words,

the platform does not impose a no-surcharge-rule as a condition for the providers to be affiliated with the system; 2) the providers and consumers cannot incur transaction costs associated with a system of double prices for each item. In fact, those limitations cannot be imposed on our platform because of the huge demand on both sides. The platform would oblige users (the providers/the consumers), once they connect to the platform, not to sell/purchase the data directly irrespective of whether or not the providers and consumers know each other via the platform. Thousands or millions of users may connect to the platform to trade the data, and thus it is not sufficient or realistic to monitor the users to know whether or not they are in compliance with their commitment (not trade directly). Consequently, signing an agreement between the users and the platform to prevent the users from direct trade does not sufficiently work. In addition, monitoring the users raises privacy issues, which imposes high costs on the platform. Alternatively, we study the efficiency of the platform in an open environment where consumers and providers can negotiate to avoid using the platform and interact directly.

2.8.3 Collecting and Sharing Data Literature

To the best of our knowledge, personal data monetization has not been studied extensively. Few proposals [52, 44, 18, 55] have addressed data marketing in terms of privacy concerns, where organizations are considered as second owners of data. The authors in [55] are inspired by economics-based approaches to disseminate sensitive data to third parties. This paper, like all the proposals in this context, focus on the relationship between privacy concerns and data marketing, without the interventions of the individuals, i.e. the actual data owner. However, the work follows the classical form of intermediaries where organizations buy the data and sell them to consumers.

The problem of collecting and sharing data in the domain of mobile phone sensing applications has been addressed using different models and techniques. [75] designed a recruitment framework identifying well-suited participants for data sensing based on spatial-temporal availability as well as participation habits. [89] designed a bidding model in order to minimize the cost of data collection for crowdsourcing applications. The model follows the form of peer-to-peer interactions between data consumers and participants and sees the problem by the consumer's eyes (rather than individuals); where the proposal focuses on the buying process with minimal costs rather than the idea of

trading data that requires different machinery. [104] have proposed two game theory-based incentive mechanisms to motivate individuals to involve in the mobile sensing tasks: A platform-centric incentive mechanism that is modeled using the Stackelberg game and a user-centric incentive mechanism that is modeled using the reverse auction game. Moreover, [53] and [43] have used auction mechanisms to motivate and reward truthful contributions. However, the nature of personal data play against auctions as a successful sale mechanism. Specifically, the same data can be resold to many data consumers; which cancels the auction concept that entails one winner of buyers. Furthermore, the enormous demand of bidders or auctioneers (data providers and data consumers) adds high cost in terms of time and complexity of the interactions, which makes auctions unpractical as a mechanism for our data monetization platform. Additionally, these proposals are complex to implement in a fully-distributed and highly-dynamic setting. [70, 48, 46] address the optimal pricing mechanisms and data management for data analytic services. Those proposals, in general, design a data market model consisting of of three entities 1) data vendor; 2) service provider; and 3) service customers. The service provider first buys the raw data from the data vendor. Then, the raw data is processed and analyzed by the service provider to develop advanced models, for example, using machine learning techniques, and to offer services to the service consumers. Those proposals have generally designed the data market using the merchant paradigm to provide information services. However, we discussed extensively the drawbacks and concerns of such model as successful model for data trading.

To the best of our knowledge, this work is the first economics-based proposal that addresses the monetization of data from the perspective of individuals and consumers and that uses the two-sided market concept. We have provided the idea of monetizing data for the first time in [17]. However, the analysis has not been mature enough and suffered from many drawbacks such as the linear demand model. This proposal differs from the old one as follows: 1) we update our analysis and replace linear demands by logarithmic demands simulating realistic behaviors of the consumers and the providers; 2) we provide a rigorous analysis for platform equilibrium over different ranges of externalities; 3) we examine the success of the platform in the unregulated case; and 4) we discuss output results in different domains of collecting and sharing data. While we have presented some proposals that have a direct link (e.g. privacy concerns and data marketing) or indirect link

(crowd sensing) with data monetization, our work differs from those proposals as shown in the following: 1) we provide a platform for monetizing data rather than incentives for providers that focus on achieving a particular level of users' participation; and 2) we use two-sided theories for data monetization. Our framework is designed so that it attracts more data providers and data consumers and, as a consequence, increases the providers' payoff, and guarantees an adequate level of data amounts.

2.9 Conclusion and Future Work

In this paper, we propose and analyze a novel platform for personal data monetization using two-sided market theory. The proposed platform is a coordinated marketplace that facilitates the search for data providers and data consumers and allow them meet to exchange financial benefits. The proposed platform provides a solution for using personal data by involving individuals in the data monetization process. Furthermore, the platform helps data consumers increase the engagement of data providers, and get easy and quick access to high quality personal data. Consequently, 1) searching costs have been cut dramatically; and 2) user's privacy has been boosted by giving individuals the control of with whom they share specific personal data and get proper compensation. The two-sided market has been investigated as a powerful economic model for such a platform. The implemented grounded theory (i.e., two-sided market) provides an efficient mechanism to attract and increase the engagement of data providers and consumers. Compared to the merchant model, as shown experimentally in the paper, the two-sided model increases the total surpluses including providers, consumers and the platform payoff.

The theoretical analysis revealed that The platform is efficient and still receives positive payoff in the following situations: 1) medium level of externalities; 2) low level of users stability; and 3) in domains that require relatively large numbers of providers. The paper recommends to use secure technologies such as blockchain to hide identities of consumers and providers and preserve their privacy in order to alleviate limitations concerning users stability. As future work, we intend to move the theoretical model to cloud computing where the biggest chunks of data and IoT services reside. We will also extend the platform to deal dynamically with changes on consumers and providers

sides. This step requires further investigation using different economic and computation techniques to consider cloud resources elasticity, particularly game theory.

Chapter 3

Cloud Computing as a Platform for Monetizing Data Services: A Two-Sided Game Business Model

With the unprecedented reliance on cloud computing as the backbone for storing today's big data, we argue in this paper that the role of the cloud should be reshaped from being a passive virtual market to become an active platform for monetizing the big data through Artificial Intelligence (AI) services. The objective is to enable the cloud to be an active platform that can help big data service providers reach a wider set of customers and cloud users (i.e., data consumers) to be exposed to a larger and richer variety of data to run their data analytic tasks. To achieve this vision, we propose a novel game theoretical model, which consists of a mix of cooperative and competitive strategies. The players of the game are the big data service providers, cloud computing platform, and cloud users. The strategies of the players are modeled using the two-sided market theory that takes into consideration the network effects among involved parties, while integrating the externalities between the cloud resources and consumer demands into the design of the game. Simulations conducted using Amazon and google clustered data show that the proposed model improves the total surplus of all the involved parties in terms of cloud resources provision and monetary profits compared to the current merchant model.

3.1 Introduction

Cloud computing is witnessing a striking increase in the number of enterprises and manufacturers that are relying on this paradigm to store and process their data. For example, the study reported in [2] revealed that one million customers deploy their own enterprises on Amazon, spending 30 billion USD on persistent storage on Amazon EC2 instances and generating 600 ZB of data per year [46]. This explosive amount of data generated and stored on cloud resources forms the backbone for Artificial Intelligence (AI) services and opens the door for a new cloud business paradigm, enabling the latter to be an active platform for monetizing data that benefit AI services. However, the cloud is not the actual owner for these big chunks of data, and has no right to trade and use these data without considering its actual owners.

Motivated by the vision of the cloud as platform for monetizing data services, we propose in this paper a novel cloud business model which allows data consumers (e.g., market research enterprises) to run their data analytics on the huge and diverse data that are stored on the cloud. This not only gives data consumers the opportunity to extract valuable patterns from massive data coming from multiple data providers, but also releases them from having to search and discover appropriate providers for each particular type of data they need to analyze. Data providers, in addition to favoring the access to cloud-based infrastructure over purchasing their own computing and storage platforms, find in the enormous and varied number of data consumers that deal with the cloud an extra motivation to store their data on this platform to improve their exposure and increase their market shares. This indirectly makes, as shown in Figure 3.1, the cloud computing platform a mediator between data providers and data consumers and a principal player in the whole big data analytics process. This opens the door for new and innovative business models to take advantage of this scenario to increase the profits of all the involved parties, apart from the traditional business models which treat the cloud as being a passive virtual market for offering services via the Internet.

Specifically, the literature on business-oriented data trading can be classified into two main categories, i.e., pure merchant approaches and collaborative approaches. The proposals under the pure merchant approach such as [70, 48] and [46] adopt classic economic approaches, mainly

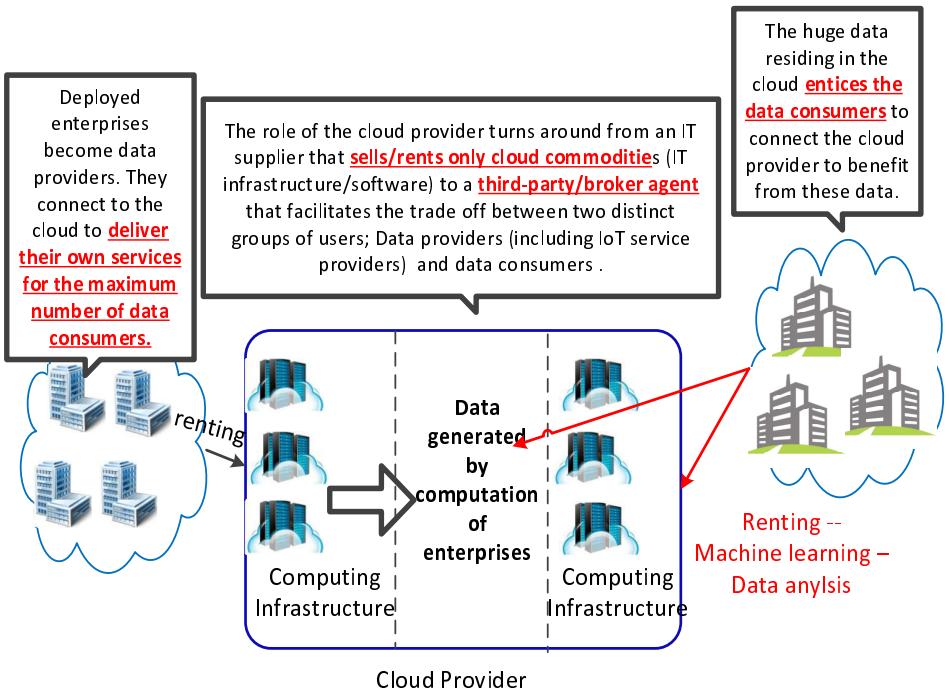


Figure 3.1: Overview of the new cloud business model

the demand-supply model and one-sided game theory/auction-based pricing to model the interactions among data providers, data consumers and third-party platforms (i.e., information service providers). In this approach, the third-party platform aims to maximize its revenue through buying data from their owners, reprocessing them, extracting useful information and selling this information to consumers. This approach suffers from several limitations when applied in cloud computing scenarios. The first limitation is related to the diversity in the data consumers' interests, which entails higher processing costs (in terms of information extraction for different customers' interests) for the third-party platform. Moreover, under the pure merchant model, data providers aim to maximize their revenue of using their data commodities while the third-party platform aims to minimize the cost of raw data bought from these providers. In parallel, from the data consumers' side, the third-party platform aims to maximize its revenue from selling the processed information while consumers aim to minimize the cost of information commodities, considering the maximum available quality and quantity of information. The resulting equilibrium from such aggressive competitions among the different involved parties leads to less and coarse distribution of the total surplus. In

addition, data differ from other economic goods for its potential of being (re)-sold to many consumers at the same time. In the pure merchant model, since economic goods cannot be resold, the equilibrium of the market lies at the intersection of the demand and supply curves. This means that the quantities of goods needed by consumers is equal to the quantities of goods provided by the sellers. This however does not hold in our case since the same data can be shared with more than one consumer at a time, which leads to an aggressive competition among data providers to sell their data even at lower prices. The drawbacks of the merchant model are deply discussed in [16] which alternatively proposes a two sided market model for monetizing personal data. In more detail, Bataineh et al.[16] propose an open market model in which individuals (actual data owners) and data consumers trade data over a third party platform that helps them discover each other. The authors show that the two-sided market outperforms the merchant model in maximizing the total surplus. However, the main limitation of this approach is that it is based on a static analysis of consumer' demand and data prices, which makes it unsuitable for dynamic cloud markets.

Under the umbrella of collaborative approaches, some proposals, for instance [23] and [101], tried to model the interactions among three entities in the domain of business-oriented IoT. In [23], the authors propose a model in which client peers are interested in sharing video content with the help of the cloud. In [101], the authors propose game theoretical models among IoT sensors, IoT service providers and data consumers. In these games, two entities (i.e., IoT sensors and IoT service providers) cooperate together in one game and then compete as one entity against data consumers. Such an approach suffers from three drawbacks: (1) it does not consider the cross-group externalities (e.g., the mutual impact of the clientele size of one party on that of the other party) among the involved parties, which makes it unable to capture the whole and more concrete and realistic picture of the three-sided economical model; (2) the cooperation and competition strategies adopted by the different players are highly impacted by the cross-group externalities which might not always lead to the best outcome for these players; and (3) it does not clarify how cooperating entities would share their earned revenues.

Adopting traditional game theory concepts (e.g., Shapley value and Nash equilibrium) to distribute the revenue that results from the cooperation among the different parties suffers from several limitations when applied in dynamic data trading scenarios over the cloud. Specifically, 1) although

such concepts might be highly efficient in scenarios wherein all the involved parties are rational, their effectiveness starts to decrease in the presence of parties that are heterogenious and prefer to deviate from the equilibrium points. For example, recent studies have revealed that only 37% of the players tend to accept the Nash equilibrium in cooperative games (interested readers can consult behavioral games and ultimatum games [39] for further details); and 2) even though the Shapley value approach fairly splits the revenues among the cooperative entities based on their contributions, it becomes inapplicable in cases wherein the contributions of entities cannot be measured (which applies to the cloud scenario considered in this work). Specifically, the cloud provider adds an ethereal/intangible, yet significant, contribution to the coalition via introducing the wide social networks of data consumers to those of data providers. On other hand, data providers own the data which forms the core of this new business. This creates a continuous dilemma between data providers and cloud providers about who makes the most significant contribution to the coalition and hence who deserves the biggest share of the revenues. Equal distribution, so-called *fifty-fifty*, is one approach to split the revenues between the cloud provider and data provider. However, as mentioned before, the rationality and greediness of the involved parties (i.e., the cloud provider and data provider) prohibit the success of such a strategy. This leads us to the conclusion that we are dealing with a behavioral and ultimatum game in which two players (proposer and responder) argue to split a certain amount of revenue. The proposer is endowed with a sum of revenue and is responsible for splitting this sum with the responder. The responder may accept or reject the sum. In the case the responder accepts the sum, the revenue is split as per the proposal; otherwise, both players receive nothing.

Contributions. To solve the aforementioned problems, the two-sided market model [82], which is praised for its success in modeling situations that involve brokers and cross-group externalities, is investigated to study the cloud-based data trading problem. The main idea of our solution is that the cloud computing platform tries to attract data consumers by offering them higher amounts of computing resources to deploy their data analytic tasks. This in turn contributes in attracting a larger number of data providers to reach the cloud's network of data consumers. Consequently, the data providers have incentives to offer higher portions of their revenues to the cloud computing platform. Two-sided market provides effective solution concepts for situations that are characterized by a

third-party platform connecting two other parties. However, the main limitation of the two-sided market theory is that it is effective in modeling scenarios in which the demand is static, but becomes less effective in elastic environments that characterize cloud computing where the demand is subject to dynamic and continuous changes. To address this problem, we integrate a novel game theoretical model, as shown in Figure 3.2, on top of the two-sided market model. The players of our game are (multiple) independent competing service providers (followers) and the cloud computing platform (leader). The players opt for hybrid cooperative and non-cooperative strategies, where strategies are modeled as closed loops of dependencies. Data consumers and the cloud platform exhibit cross group externalities between each other, where a higher demand from consumers leads to a revenue increase for the cloud platform and a higher supply of computing resources from the cloud creates more demand from consumers.

In the first stage of the game, the leader (cloud platform) announces the desired portion of returned revenues out of the data providers' gain, and then in the second stage, data providers decide about their pricing strategy for data consumers. The resulting equilibrium forces the cloud platform to offer higher and reasonable supply of computing resources to guarantee maximal levels of revenues, while not showing greedy behavior in terms of its share of data providers' revenue. Moreover, following our solution, the data providers are forced to offer the cloud platform a higher portion of their revenues to ensure appropriate Quality of Service (QoS) delivered to data consumers. In the case of a greedy behavior from the cloud, our game uses a subsidizing mechanism. This mechanism pushes data providers to increase the shared portion offered for the cloud to sustain high and reasonable levels of computing infrastructure so as to guarantee high levels of consumers' demand. Similarly, in cases where data providers behave greedily by offering small portions of revenues to the cloud, the subsidizing mechanism pushes the cloud to pump out more infrastructure units to increase the consumers' demand so as to guarantee the highest possible level of revenue portion.

To validate our solution, we conduct empirical experiments using real-world data from Google and Amazon. Experimental results show that by following our solution, all the involved parties (i.e., cloud platform, data providers and data consumers) achieve higher revenues than those achieved by the traditional cloud computing business model.

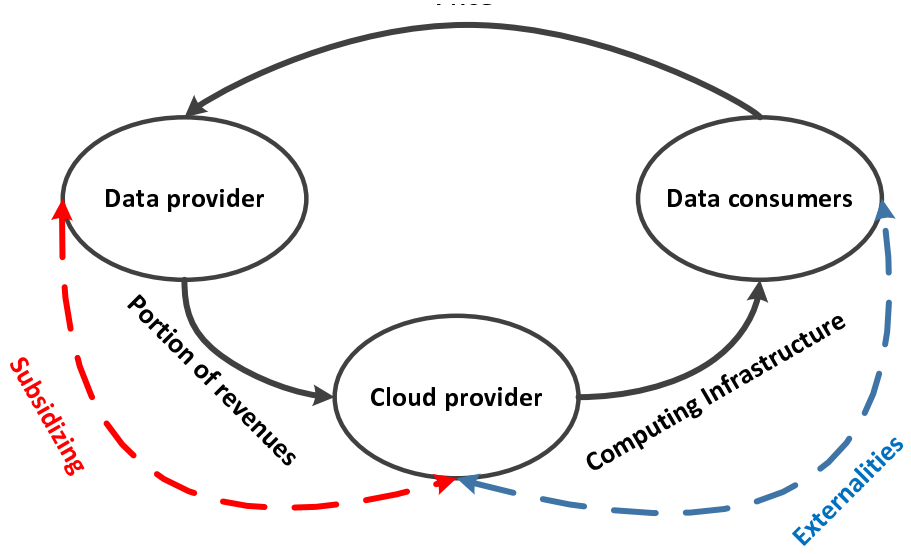


Figure 3.2: Overview of the proposed two-sided game

3.2 Related Work

In this section, we provide a literature review on cloud computing business models. The existing proposals can be classified into two main categories: classical market and game theoretic-based pricing models. The proposals under the classical market category such as [38, 109, 74] tackle the pricing of the cloud services using simple pricing models including those of cost-based pricing, differential pricing, Ramsey pricing, and demand curve function. They model the pricing of cloud computing resources as an optimization problem among multiple cloud providers and cloud users. However, the main drawback of these approaches lies in their static pricing strategy which does not suit the highly variable and dynamic environment of cloud computing.

On the other hand, the game theoretical models consider the instantaneous interactions that might occur among the involved entities and their effects on each party's welfare. The objective is to dynamically capture the optimal price and distribution of the cloud computing resources. Many proposals such as [57, 71, 103, 33, 88, 8] applied different approaches including games and machine learning [97, 77, 81] to the cloud resource allocation and pricing problem. In [57], the authors propose an economic model based on a Stackelberg game to trade video contents and movies over

a cloud platform. The proposed model formulates the interactions between a service provider (e.g., Netflix) and end users. The service provider acts as the game leader and aims to minimize the cloud bandwidth consumption while guaranteeing at the same time users' satisfaction. The work in [95] models the interactions among multiple Software as a Service (SaaS) providers and Infrastructure as a Service (IaaS) provider as a two-stage Stackelberg game. In the first stage of the game, SaaS providers determine the number of required VM instances while accounting for both the QoS delivered to their users and the associated costs. In the second stage, the IaaS providers seek to maximize their revenues in the light of the bids done by the SaaS providers [60, 59]. The author in [23] proposes an economic model in which cloud users seek to share video content with other users over the cloud. The model is solved using both cooperative and non-cooperative games between the cloud and its users. Similar studies are investigated in [101, 33] for different cloud applications. The authors in [110] propose a game theoretical model to deliver a bundle of complementary IoT services. The proposed solution studies the merchant-consumer scenario in which the IoT services are directly traded between the service providers and service consumers without the intervention from any third party. However, this solution cannot be adopted in our case, where the cloud computing is not the actual data owner and hence it cannot monetize the data directly for the consumers. Nevertheless, the cloud computing (the third party in our paper) is considered as a global market where the data services and data consumers meet each other, thus increasing their market shares. The authors in [17] and [62] introduce a market model for managing, trading, and pricing big data services. Both proposals use the two-sided market theory in order to provide incentives for both cloud providers and users to increase their data shares. The work presented in [16] extends the work proposed in [17] and comprehensively studies the two-sided market model as a successful model for monetizing personal data. However, these proposals consider a static environment in which the demands on cloud resources are computed in a static manner, which makes them unable to accommodate the cloud's elasticity property.

To the best of our knowledge, the proposed work is the first that addresses big data services monetization, while considering the cross-group externalities among the involved entities. Unlike the classical cloud computing business model (where the main challenge is how to optimize the cloud utilization while incorporating only operational cost and QoS metrics), our approach : 1)

supports and helps junior big data service providers especially those that have limited monetary budgets; 2) uses the two-sided market theory to model the interactions among the involved parties, while all above-discussed proposals use the classical merchant model; (3) includes a subsidizing technique to push the resulting equilibrium toward a Pareto optimal point. On the other hand, the above-discussed proposals adopt the fairness criterion that rewards the involved parties based on their contributions. We also differ from the other proposals that adopt the two-sided market theory by providing a dynamic pricing method, instead of a static game theoretic-based pricing strategy.

3.3 Proposed Big Data Services Monetization Model over the Cloud: A Two-sided Game Model

We explain in this section the details of our proposed Big data services monetization.

3.3.1 Solution Architecture and Game Formulation

The proposed cloud market platform, depicted in Figure 5.3, consists of three entities: consumers of services CS (CS_i denotes Consumers of Service i), big data service providers SP (a Service Provider providing service i is denoted SP_i) and a typical Cloud Platform (CP). The cloud platform, such as Google and Amazon, is a market leader with huge computing and storage capabilities, capitals, and social consumer networks. In our model, a certain big data service provider SP_i that provides a service i deploys its service on the cloud and receives a monetary value of P_i for each consumer access to its service i. The cloud platform CP is in charge of sustaining the consumer access through providing the needed computing and storage infrastructure including hardware, software and security services. The relationship between consumer CS_i's demand, denoted by D_{c_i}, and the computing and storage resources D_{s_i} supplied to CS_i is modeled using the two-sided market model as cross group externalities α and β. Here, α represents the benefits that a consumer obtains when some new computing and storage resources are added to D_{s_i} and β represents the amount of benefits that the cloud platform earns when more new consumers are added to D_{c_i}. The parameters α and β are dependant on the service i. However, instead of using the notations α_i and β_i, the index i is omitted to simplify the equations where the service i is understood from the context. The same simplification is used for the other parameters that appear as powers (exponents) in our equations.

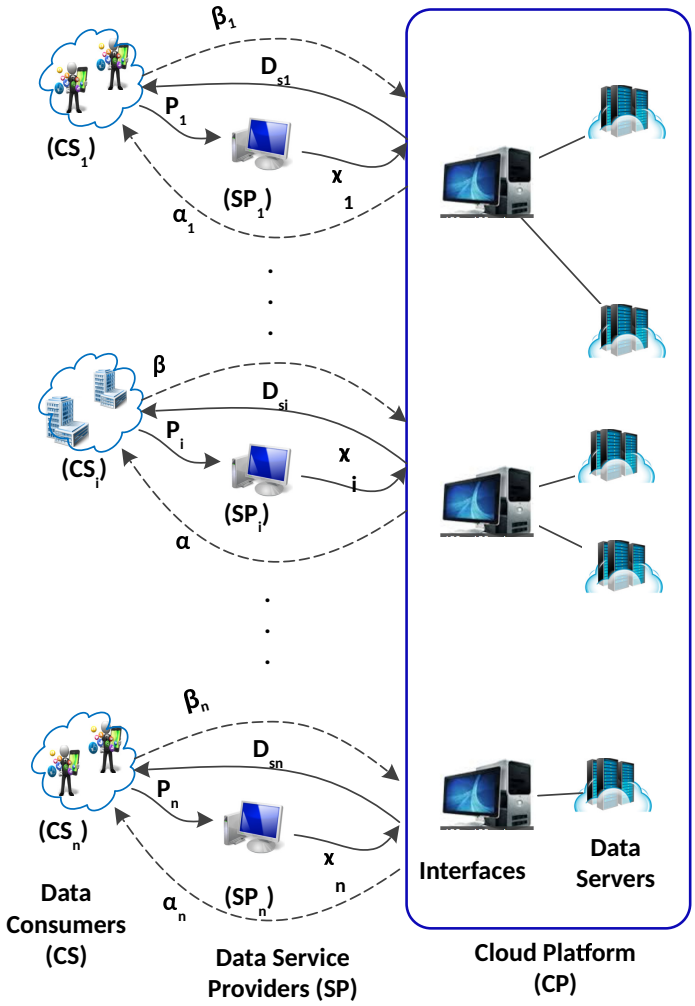


Figure 3.3: Two-sided model

The interaction between SP and CP is modeled as a two-stage game where CP acts as the game leader and SP are the followers. In the first stage of the game, each service provider providing service i SP_i observes the amount of money returns χ_i requested by CP, in order to adjust the price to be charged to CS_i. In quest of the price specified by SP_i, CP determines the optimal amount of computing and storage resources D_{s_i} that should be supplied to CS_i. The model forms a closed loop of dependencies that involves techniques from Stackelberg and Ultimatum game theory as well as a subsidizing technique.

In the Stackelberg game, the interactions take place in two stages where the leader (CP) makes the first move and then does each follower (SP_i) after having observed the leader's move. In the

ultimatum game, the first player (CP) proposes a strategy to divide the amount of returned revenue with the second player (SP_i). In case SP_i rejects the offer, neither player gains anything. Otherwise, the first player gets the amount it requested and the second player gets the rest. In the subsidizing technique, SP_i may chose to subsidize CP by an extra amount of payment that exceeds the contribution of this CP. The objective is to keep an optimal level of D_{s_i} that maximizes the return revenues $P_i * D_{c_i}$. Alternatively, CP may subsidize SP_i by low portion of the resulting revenues to keep an optimal level of P_i. The different parameters and symbols used in our proposed solution are depicted in Table 3.1.

Model Parameters	Descriptions.
SP_i	Service provider providing service i.
CP	A typical cloud platform.
CS_i	Consumers of service i.
D_{c_i}	CS_i's demand.
D_{s_i}	IT-infrastructure supply to CS_i.
P_i	Service i's price.
ϕ	D_{s_i}'s elasticity with respect to χ_i.
χ_i	Portion of revenue required by CP from SP_i.
γ	D_{c_i}'s elasticity with respect to P_i.
β	The Network effects (externality) on D_{s_i} by D_{c_i}
ψ	D_{s_i}'s elasticity with respect to P_i.
α	The Network effects (externality) on D_{c_i} by D_{s_i}.
f_c	Associated costs per service consumer.
k_1, k_2	Constant multipliers.
π_i	SP_i's payoff.
f_s	Associated costs per IT-infrastructure unit.
π	Cloud platform's payoff.
a_1	$= \gamma - \alpha\psi$.
a_2	$= 1 - \alpha\beta$.
a_3	$= \psi - \gamma\beta$.
a_4	$= \alpha\phi$

Table 3.1: Model parameters

3.3.2 Players' Demand and Utility Functions

A precise estimation of the needed computing and storage resources requires a price estimation mechanism for the number of consumers and the variation of their demand with respect to the provided QoS. To do so, we define the consumer's demand and supply using the Cobb-Douglas

function that effectively captures the elasticity of the computing and storage resources supply (D_{s_i}) and its variations for each specific user's demand. This elasticity is a characteristic property of cloud computing environments. The demand functions we use are continuous, concave or convex, and capture the elasticity with respect to each input parameter. Two elasticity parameters are used γ and ψ (see Table 3.1). These two parameters depend on the service i, which is omitted from the notations for simplicity as mentioned earlier. In our model, the consumer's demand (D_{c_i}) is a function of P_i and D_{s_i} as shown in Equation (23).

$$D_{c_i} = k_1 P_i^{-\gamma} D_{s_i}^{\alpha} \tag{23}$$

D_{s_i} is given in Equation (24). Clearly, higher consumers' demands would have an influence on the quantity of supplied resources. The cloud platform CP uses more computing and storage resources to keep up with the increasing number of consumer accesses, to maintain a high quality level. The parameter χ_i^{ϕ} represents the cloud platform's preferences (i.e., desired profit) and implicitly captures the rationality of both CP and SP_i. In fact, it reflects the level of perfect/imperfect information that CP and SP_i have about one another. High elasticity ϕ is caused either by a greedy monopolist cloud platform or by a weak service with few capitals accepting small portions of returns on profits. The parameter ϕ depends on the service i, but as mentioned earlier, the index i is omitted when the service i is understood from the context. The charged price P_i also positively contributes to D_{s_i}. We can arguably claim that charging consumers with higher prices P_i forces CP to provide more computing and storage resources so as to satisfy the consumers' needs. Modeling D_{s_i} as a function of χ_i and P_i with different elasticity values connects CP and SP_i strategies with each other, which captures the sensitivity of CP to SP_i's strategy (i.e., structure of the charged price and shared portion), and highlights the importance of the subsidizing technique. This aspect is illustrated and discussed further in the simulation section (Section 3.4.4).

$$D_{s_i} = k_2 \chi_i^{\phi} P_i^{\psi} D_{c_i}^{\beta} \tag{24}$$

By substituting Equation (24) into Equation (23) and vice versa, we can express D_{c_i} and D_{s_i} as functions of P_i and χ_i as follows:

$$D_{c_i} = (k_1 k_2^{\alpha} P_i^{-a_1} \chi_i^{a_4})^{1/a_2} \tag{25}$$

$$D_{s_i} = (k_2 k_1^{\beta} P_i^{a_3} \chi_i^{\phi})^{1/a_2} \tag{26}$$

Each big data service provider SP_i is subject to a fixed cost f_c per each consumer access. SP_i aims to maximize its payoff as given in Equation (74). We express the service provider's payoff π_i as a function of P_i and χ_i by substituting Equation (25) into Equation (74) and taking the log for both sides as shown in Equation (78).

$$\pi_i = ((P_i)(1 - \chi_i) - f_c) D_{c_i} \tag{27}$$

$$\begin{aligned} \log \pi_i = \log(P_i(1 - \chi_i) - f_c) + (1/a_2)(\log k_1 k_2^{\alpha} \\ - a_1 \log P_i + a_4 \log \chi_i) \end{aligned} \tag{28}$$

The cloud platform CP is subject to a fixed cost f_s per each unit of computing and storage resources. The CP aims to maximize its payoff as given in Equation (29). We express the cloud platform's payoff π as a function of P_i and χ_i by substituting Equations (25) and (26) into Equation (29) as shown in Equation (30).

$$\pi = P_i \chi_i D_{c_i} - f_s D_{s_i} \tag{29}$$

$$\pi = (k_1 k_2^{\alpha})^{\frac{1}{a_2}} P_i^{1 - \frac{a_1}{a_2}} \chi_i^{\frac{a_4}{a_2} + 1} - f_s((k_2 k_1^{\beta})^{\frac{1}{a_2}} P_i^{\frac{a_3}{a_2}} \chi_i^{\frac{\phi}{a_2}}) \tag{30}$$

3.3.3 Game Equilibrium

The equilibrium of the above-described game is solved using a backward induction methodology. Thus, the followers' (service providers) sub-game is solved first to obtain their response P_i to the service consumers. The leader's (cloud platform) sub-game is then computed considering all the possible reactions of its followers to maximize its payoff [96]. Every service provider SP_i

determines its optimal decision P_i^*, while considering the CP's optimal decision χ_i^* as an input parameter. The players' best responses are discussed in the following.

Theorem 2. *The best responses in the two-sided game are as follows:*

(1) The best response of the service provider SP_i is given by:

$$P_i^* = \frac{a_1 f_c}{(a_1 - a_2)(1 - \chi_i)} \tag{31}$$

if: $\frac{a_1}{a_1 - a_2} > 0$ *and* $\frac{a_1}{a_2} > 1$

(2) The best response of the cloud platform with respect to a service i is given by:

$$\begin{aligned} \chi_i^{\frac{a_4-\phi}{a_2}+1}(1-\chi_i)^{\frac{a_1+a_3}{a_2}-1} &= f_s \times (\frac{\phi}{a_4+a_2}) \\ &\times (\frac{k_2 k_1^{\beta}}{k_1 k_2^{\alpha}})^{\frac{1}{a_2}} \times (\frac{a_1 f_c}{a_1 - a_2})^{\frac{a_1+a_3}{a_2}-1} \end{aligned} \tag{32}$$

if: $a_4 + a_2 - \phi < 0$

Proof. Consider the service payoff given by Equation (78), the optimal price P_i^* is defined by $\partial \pi_i / \partial P_i = 0$ as follows:

$$\frac{1}{\pi_i} \times \frac{\partial \pi_i}{\partial P_i} = \frac{1 - \chi_i}{P_i(1 - \chi_i) - f_c} - \frac{a_1}{(a_2) P_i} = 0 \tag{33}$$

$\Rightarrow$

$$P_i^* = \frac{a_1 f_c}{(a_1 - a_2)(1 - \chi_i)} \tag{34}$$

Since P_i^* is always positive, then

$$\frac{a_1}{a_1 - a_2} > 0 \tag{35}$$

To verify the type of P_i^*'s optimality, i.e maximum or minimum, we compute a second derivative test by deriving Equation (74):

$$\frac{\partial \pi_i}{\partial P_i} = (1 - \chi_i) D_{c_i} + (P_i(1 - \chi_i) - f_c) \frac{\partial D_{c_i}}{\partial P_i} \tag{36}$$

By deriving Equation 25, then

$$\frac{\partial D_{c_i}}{\partial P_i} = \frac{-a_1}{a_2 P_i} D_{c_i} \tag{37}$$

By substituting Equation 37 into Equation 36, then

$$\frac{\partial \pi_i}{\partial P_i} = (1-\chi_i)D_{c_i} - \frac{a_1}{a_2 P_i}(P_i(1-\chi_i) - f_c)D_{c_i} \tag{38}$$

By rewriting Equation (38) using Equation (74), then

$$\frac{\partial \pi_i}{\partial P_i} = (1-\chi_i)D_{c_i} - \frac{a_1}{a_2 P_i}\pi_i \tag{39}$$

$$\frac{\partial^2 \pi_i}{\partial P_i^2} = \frac{-(1-\chi_i)a_1}{a_2 P_i}D_{c_i} - \frac{a_1}{a_2 P_i}\frac{\partial \pi_i}{\partial P_i} + \frac{a_1}{a_2 P_i^2}\pi_i \tag{40}$$

By simplifying Equation (40) and substituting Equation (34), we obtain:

$$\frac{\partial^2 \pi_i}{\partial P_i^2} = \frac{D_{c_i}}{P_i}(1 - \frac{a_1}{a_2})(1-\chi_i) \tag{41}$$

Since D_{c_i} and P_i are always positives, then

$$\frac{\partial^2 \pi_i}{\partial P_i^2} < 0 \Rightarrow (1 - \frac{a_1}{a_2}) < 0 \Rightarrow \frac{a_1}{a_2} > 1 \tag{42}$$

Similarly, to obtain the optimal χ_i^*, we derive Equation (30) with respect to χ_i as given by Equation (43):

$$\frac{\partial \pi}{\partial \chi_i} = (k_1 k_2^{\alpha})^{\frac{1}{a_2}}(\frac{a_4}{a_2}+1)P_i^{1-\frac{a_1}{a_2}}\chi_i^{\frac{a_4}{a_2}} - (\frac{\phi f_s (k_2 k_1^{\beta})^{\frac{1}{a_2}}}{a_2})P_i^{\frac{a_3}{a_2}}\chi_i^{\frac{\phi}{a_2}-1} = 0 \tag{43}$$

$$\chi_i^{\frac{a_4-\phi}{a_2}+1} = f_s(\frac{\phi}{a_4+a_2})(\frac{k_2 k_1^{\beta}}{k_1 k_2^{\alpha}})^{\frac{1}{a_2}} P_i^{\frac{a_1+a_3}{a_2}-1} \tag{44}$$

By substituting Equation (34) in Equation (44), we get:

$$\chi_i^{\frac{a_4-\phi}{a_2}+1}(1-\chi_i)^{\frac{a_1+a_3}{a_2}-1} = f_s(\frac{\phi}{a_4+a_2})(\frac{k_2 k_1^{\beta}}{k_1 k_2^{\alpha}})^{\frac{1}{a_2}}(\frac{a_1 f_c}{a_1-a_2})^{\frac{a_1+a_3}{a_2}-1} \tag{45}$$

To verify the type of χ_i^*'s optimality, we compute a second derivative test by deriving Equation (43) as given by Equation (46):

$$\begin{aligned}\frac{\partial^2 \pi}{\partial \chi_i^2} &= (k_1 k_2^{\alpha})^{\frac{1}{a_2}}(\frac{a_4(a_4+a_2)}{a_2^2})P_i^{\frac{a_2-a_1}{a_2}}\chi_i^{\frac{a_4}{a_2}-1} \\ &\quad - (\frac{f_s\phi(\phi-a_2)(k_2 k_1^{\beta})^{\frac{1}{a_2}}}{a_2^2})P_i^{\frac{a_3}{a_2}}\chi_i^{\frac{\phi}{a_2}-2} < 0\end{aligned} \tag{46}$$

By substituting Equation (44) in Equation (46), we obtain:

$$a_4 + a_2 - \phi < 0 \tag{47}$$

□

3.4 Simulations and Empirical Analysis

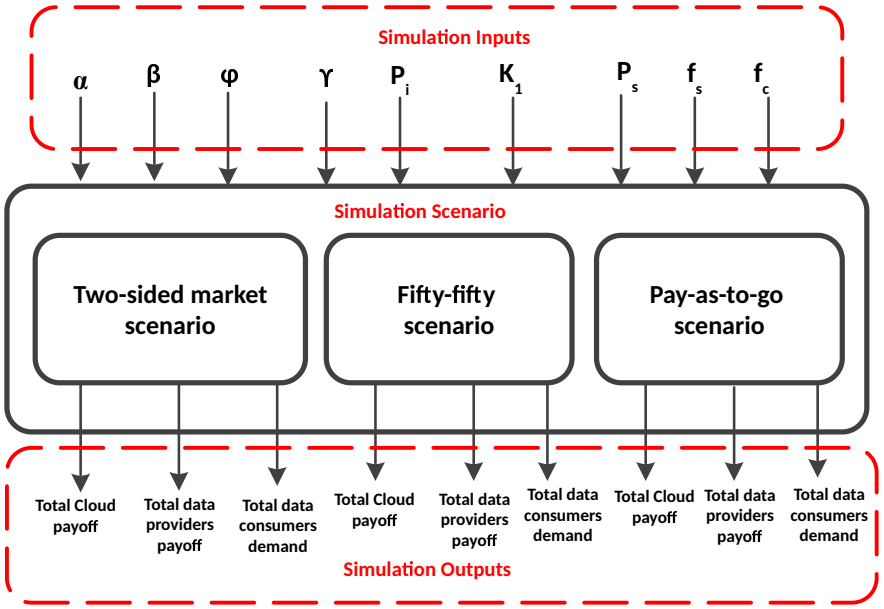


Figure 3.4: Simulation overview

In this section, we evaluate the performance of the proposed two-sided game solution in comparison with the fifty-fifty and the pay-as-to-go approaches in terms of total surpluses of involved parties (i.e., payoffs of the cloud platform, service data providers and service data consumers). Specifically, we aim to: 1) verify the effectiveness of the proposed game vis-à-vis the current cloud computing business model (i.e., the pay-as-to-go model); 2) study the equilibrium of the two-sided game in presence of the fifty-fifty choice (i.e., egalitarian choice in ultimatum games) which is the typical solution for such games; and 3) investigate the impact of the model parameters on the performance of our solution. Figure 3.4 shows an overview of our simulation setting in terms of inputs, scenarios, and results. The simulation inputs and scenarios are described in Sections 3.4.1 and 3.4.2 respectively, while the simulation results are discussed in detail in Sections 3.4.3, 3.4.4 and 3.4.5.

3.4.1 Simulation Setup

In this section, we conduct a simulation analysis grounded on statistical observations from BMR [2] and real data from [37]. According to [2], in 2019, Amazon Web services (AWS) received 30 billion USD in revenue with net income around 10 billion USD from 1 million active customers running monthly 70 million hours of their enterprises on custom instances of Elastic Compute Cloud (EC2). So, 1) an enterprise customer spends on average up to $30,000$ USD per year in monthly renting 70 hours of cloud resources; and 2) the marginal operating costs for the cloud platform is 66% of revenues (amazon received $10,000$ USD as net payoff from each consumer). By entering these numbers in the Amazon calculator [10], we can conclude that the customer rents on average 70 hours monthly of 32 instances of Amazon EC2 where each instance includes 16 VMs, 30 GB of Memory, and 1000 GB of hard disk storage at rate 36 USD/hour. The price rate (36 USD/hour) is denoted by P_s, which will be used later to calculate the cloud and data providers payoffs in the pay-as-to-go model as explained in Section 3.4.2. The cloud provider (amazon) entails 66% of instances price (36 USD/hour) as operating costs, which is 23.7 USD/hour. The operating costs are denoted in our model by f_s. In fact, 40% of revenues as a profit and 60% as an operation cost are common in business. Thus, we assume the marginal cost of data consumers (f_c) entailed by data providers has the same distribution as (f_s). The enterprise customer and its consumption of EC2 instances are represented by SP_i (service data provider) and D_{s_i} respectively. The mean of the supply function D_{s_i} consists of 32 EC2 instances. However, enterprise customers have varying business types and hence vary in terms of the amount of needed cloud resources. To model this variation in our simulations, the customers' demand on EC2 instances is normally distributed around the mean with a standard deviation of 10. This means that the co-domain of the supply function D_{s_i} ranges from 1 to 53 EC2 instances. The real dataset [37] registers the log file of computational big data jobs executed by tremendous enterprise customers over similar instances of EC2. This dataset helps us extract reliable ranges of consumers' demands D_{c_i} as well as the externalities α and β as described in what follows. The computational power of each instance, extracted from the same dataset [37], is normally distributed with a mean of 0.38 job per second and a standard deviation of 0.1. The average computational power is represented in the proposed model by the externality

factor α, which means that α ranges from 0.1 to 0.7. According to the assumption presented in [82], the cross group externalities factor should be bounded by 0 and 1, i.e., $0 < \alpha\beta < 1$. Hence, the externality factor β would range from 0 to $1/\alpha$. A consumer's demand D_{c_i} on each enterprise ranges from 0.1×1 to 0.7×53, which is 0.1 to 37 requests per second. The service price, P_i, is estimated through observing the prices of 150 business intelligence computing services including big data and IoT services located in the the AWS marketplace [9]. According to the observed prices, P_i is normally distributed with a mean of 1.7 USD/hour and a standard deviation of 0.5 USD. This means that the service prices range from 0.2 USD/hour to 3.2 USD/hour. The parameter ϕ represents the greediness of the cloud platform with respect to the service providers. The subsidizing factor $0 < \phi < 1$ represents the rational behavior (subsidizing behavior) of the cloud, while $1 < \phi$ represents the greedy behavior of the cloud platform. The price elasticities are set up between $0.1 - 0.35$ (i.e., γ), which are similar to the sensitivity of mobile/telecommunication services price shown in the literature [31]. We assume that $k_2 = 1$ in our simulation. By substituting the expected values of α, D_{s_i}, D_{c_i}, γ, and P_i into Equation (23) and considering the assumption $(k_2 = 1)$, we find that the multiplier k_1 ranges from 0.1 to 0.99. It is worth mentioning that consumers demand (D_{c_i}) and cloud resources (i.e., computing and storage resources) D_{s_i} supplied to CS_i are only estimated under simulation setup to extract a suitable range for the multiplier k_1, but they are not given as simulation inputs. The simulation calculates the expected consumer demand and the optimal supply of cloud resources as explained in Section 3.4.2. The values of all associated parameters are summarized in Table 3.2.

System Parameters	Values
P_i	0.2 - 3.2 USD per hour
ϕ	0 - 5
α	0.1 - 0.7
β	0 - $1/\alpha$
γ	0 - 0.35
P_s	36 USD per hour
k_1	$0.1 - 0.9$

Table 3.2: Simulation parameters values

3.4.2 Simulation Scenarios

We consider a group of 300 data service providers in the cloud under three scenarios: 1) proposed two-sided game; 2) fifty-fifty scenario which follows our model except that the cloud platform and data provider agree to share the revenue equality; and 3) pay-as-to-go scenario, which is the current business model adopted by the main cloud providers such as Amazon and Google.

Two-sided scenario

The two-sided model, explained in details in Section 3.3.1, is described in Algorithm 1. Given a data service price P_i, the cloud platform determines the optimal portion of revenue χ_i and the required amount of cloud resources that maximize its payoff.

Algorithm 1 Two-sided scenario

Input: α, β, ϕ,γ, k_1, f_s, f_c
Output: *CloudPayoff*, *ProvidersPayoff*, *ConsumersDemand*

1: **for each** $Data\ service\ provider\ SP_i$ **do**
2: SP_i declares its price P_i
3: χ_i is calculated through maximizing Equation (74)
4: The cloud calculates D_{s_i} through maximizing Equation (29)
5: Equation (23) is used to determine *ConsumersDemand*
6: Equation (74) is used to determine *ProvidersPayoff*
7: Equation (29) is used to determine *CloudPayoff*
8: **end for**

Fifty-fifty scenario

The egalitarian (fifty-fifty) scenario follows the two-sided market model in terms of consumer demand and supply function, thus considering the externalities among the involved parties (i.e., Equations (23) and (24)). Thus, the utilities of the cloud platform and data providers are formalized using the same payoff equations used in the two-sided model (i.e., Equations (29) and (74)). However, under this scenario, the cloud platform requests 50% of the revenue and hence the subsidizing factor ϕ is reset by 1. The fifty-fifty scenario is described in Algorithm 2. Specifically, given a shared portion $\chi_i = 0.5$ and a subsidizing factor $\phi = 1$, the data provider determines its optimal service price P_i through maximizing its payoff given in Equation (74). Thereafter, the

cloud platform calculates the optimal computing infrastructure D_{s_i} by maximizing its payoff given in Equation (29).

Algorithm 2 Fifty-fifty scenario

Input: α, β,γ, k_1, f_s, f_c
Output: *CloudPayoff*, *ProvidersPayoff*, *ConsumersDemand*

1: **for each** $Data\ service\ provider\ SP_i$ **do**
2: $\chi_i \leftarrow 0.5$
3: $\phi \leftarrow 1$
4: SP_i declares P_i through maximizing Equation (74)
5: The cloud calculates D_{s_i} through maximizing Equation (29)
6: Equation (23) is used to determine *ConsumersDemand*
7: Equation (74) is used to determine *ProvidersPayoff*
8: Equation (29) is used to determine *CloudPayoff*
9: **end for**

Pay-as-to-go model

Figure 3.5 depicts the pay-as-to-go model. As shown in the figure, the data provider rents the cloud computing infrastructure (D_{s_i}) for a price of P_s USD/hour. Thereafter, the data provider delivers its own data services to the data consumers for a price of P_i USD/hour. The data provider and cloud payoffs under the pay-as-to-go model are given in Equations (48) and (49) respectively. The consumer's demand function under the pay-as-to-go model has the same characteristics as under the two-sided market model. This means that the consumer's demand is formalized as given in Equation (23). However, the provided cloud computing infrastructure (D_{s_i}) is not given as a function under the pay-as-to-go model since the cloud platform and data consumers do not directly interact nor do they exhibit mutual cross group externalities between each other. The pay-as-to-go model is described in Algorithm 3. Given a price rate of P_s USD/hour, the data provider determines the optimal amount of rented cloud computing infrastructure through maximizing Equation (48). The optimal amount of rented cloud computing infrastructure $D^*_{s_i}$ is given in Equation (50) through substituting Equation (23) with Equation (48) and computing its derivatives with respect to D_{s_i}.

$$\pi_i = (P_i - f_c)D_{c_i} - P_s D_{s_i} \tag{48}$$

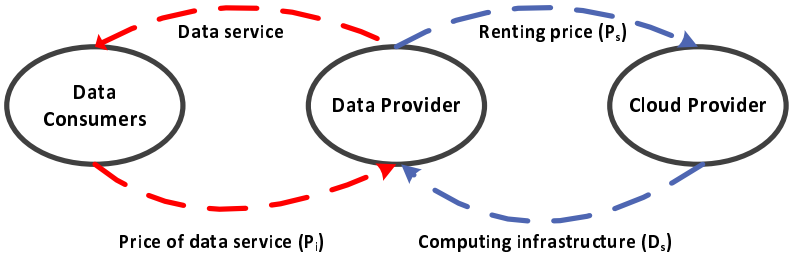


Figure 3.5: Pay-as-to-go model

$$\pi = (P_s - f_s)D_{s_i} \tag{49}$$

$$D^*_{s_i} = (\frac{P_s}{\alpha k_1 (P_i - fc) P_i^{-\gamma}})^{\frac{1}{\alpha - 1}} \tag{50}$$

Algorithm 3 Pay-as-to-go scenario

Input: α, β,γ, P_s, k_1, f_s, f_c
Output: *CloudPayoff*, *ProvidersPayoff*, *ConsumersDemand*

1: **for each** $Data\ service\ provider\ SP_i$ **do**
2: The cloud declares P_s
3: SP_i declares P_i
4: SP_i calculates D_{s_i} using Equation (50)
5: Equation (23) is used to determine *ConsumersDemand*
6: Equation (48) is used to determine *ProvidersPayoff*
7: Equation (49) is used to determine *CloudPayoff*
8: **end for**

3.4.3 Sensitivity Analysis of Externalities

We first investigate in Figures 3.6, 3.7, and 3.8 the impact of the externality parameters $(\alpha\beta)$ on the payoffs of the cloud platform, data providers, and data consumers respectively. We run the simulation with different ranges of the subsidizing factor (ϕ), i.e., 0.5, 1.0, 1.5, 2.0, and 5.0. Those values of externality and subsidizing factor are given as inputs to the simulation program to adjust the strategies of the involved players. Specifically, the cloud platform adjusts the amount of provided infrastructure (D_{s_i}) and the demanded portion (χ_i), while the data provider (SP_i)

calculates the impact of variation in (D_{s_i}) and (χ_i) on the expected demand of data consumers (D_{c_i}) and accordingly adjusts its price (P_i).

In Figure 3.6 , we study the impact of varying the externality parameters $(\alpha\beta)$ and subsidizing factor (ϕ) on the cloud's payoff. As shown in the figure, the cloud platform obtains in general higher payoff when it follows the two-sided model, rather than the pay-as-to-go model. For example, under the two-sided model, the cloud platform receives 1200, 500 and 400 USD as payoff when $(\alpha\beta) = 0.2$ and $\phi = 5.0$, 2.0 and 1.5 respectively. On the other hand, under the pay-as-to-go model and under the same externality and subsidizing parameters, the cloud platform receives less payoff of (200) USD. Similarly, data providers receive higher payoff and data consumers' demand is increased under the two-sided model compared to the pay-as-you-go model as shown respectively in Figures 3.7 and 3.8 .

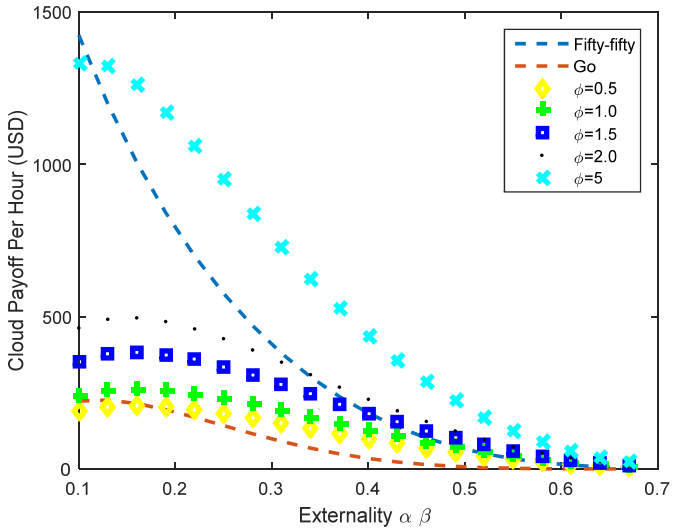


Figure 3.6: Cloud payoff over externalities $\alpha\beta$

In addition, we study the impact of the fifty-fity (egalitarian) choice on the two-sided market model, where the cloud platform and data providers share fifty percent of revenues considering the sitting of our two-sided model. As shown in Figure 3.6, the fifty-fifty choice shows more efficient outcomes than the two-sided model in terms of cloud payoff under week externalities $(\alpha\beta \in [0.1 - 0.3])$ and certain values of subsidizing factor ϕ. However, the cloud platform can receive higher payoff if it chooses higher subsidizing factor ϕ such as $\phi = 5$. Nevertheless, the egalitarian choice

shows less efficient outcomes when the externalities become stronger (i.e $\alpha\beta \in [0.3 - 0.7]$) where the cloud platform receives less payoff under certain subsidizing factor such as $(\phi = 1.5)$ or $(\phi = 2)$. On the other hand, the fifty-fifty choice and the two-sided model show similar outcomes in terms of data providers' payoff and consumers demand as shown in Figures 3.7 and 3.8 respectively. However, as mentioned in Section 3.1, the involved parties do not follow the egalitarian choice despite its good outcomes in some cases, mainly because of their greediness and because of high subsidies in real cloud markets.

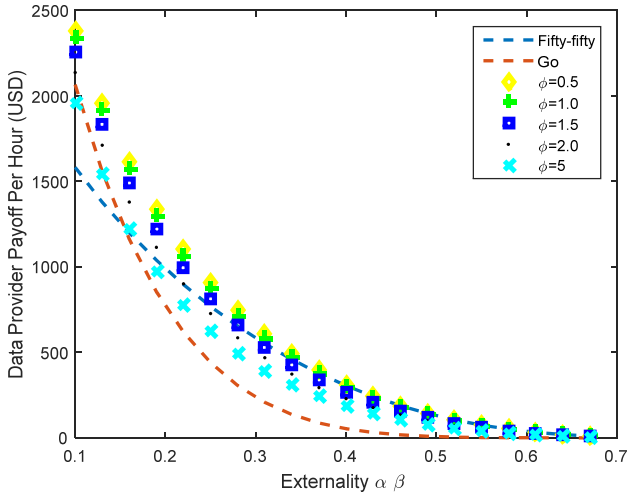


Figure 3.7: Data providers payoff over externalities $\alpha\beta$

As can be observed from these figures, the two-sided market model shows less efficient outcomes as the externalities become stronger. Specifically, the cloud platform, data providers and consumers' surpluses gradually decrease as the values of the externality parameters increase. For example, in Figure 3.7, the data providers receive a payoff of 800 USD at an externality parameters value of $(\alpha\beta = 0.2)$ and subsidizing factor of $\phi = 5.0$, which is higher than that received with externality parameter value of $(\alpha\beta = 0.3)$ and subsidizing factor of $\phi = 5.0$. This decrease continues until the two-sided market model reaches almost the same efficiency of the pay-to-go model under the strong externality values of [0.55 - 0.7]. Similar behavior is observed in terms of cloud payoff and data consumers' demand as depicted in Figures 3.6 and 3.8 respectively. In fact, the cloud's payoff and data consumers' demand are higher under weak externality values, i.e., $[0 - 0.5]$ rather than in strong externality values $[0.5 - 0.7]$. The reason behind such a behavior is that the

cloud platform needs to provide more computing and storage resources under strong externalities to attract smaller number of data consumers, which adds more costs and then leads to less payoff.

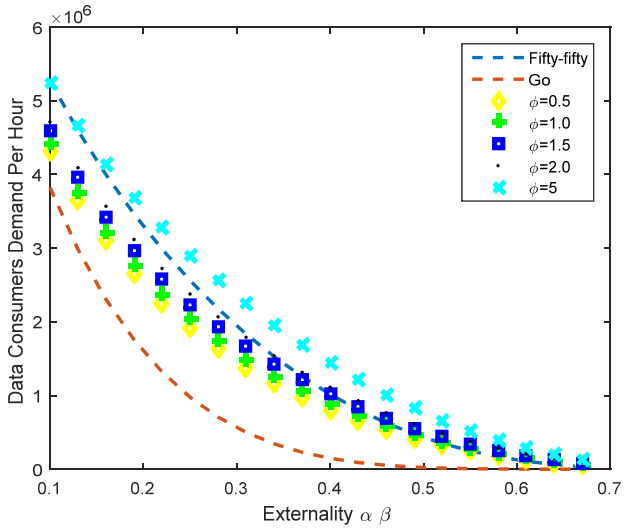


Figure 3.8: Data consumers demand over externalities $\alpha\beta$

Figure 3.9 shows the number of computing infrastructure units provided by the cloud platform over varying externality values. As shown in the figure, at a subsidizing factor of $\phi = 2.0$ and externality values of $(\alpha\beta = 0.2)$ and $(\alpha\beta = 0.53)$, the cloud platform provides 4800 units of computing infrastructure. However, in Figure 3.8, at a subsidizing factor of $(\phi = 2.0)$ and externality values of $(\alpha\beta = 0.2)$, the cloud platform attracts (3.3×10^6) data consumers, while it attracts a smaller number of data consumers, i.e., (0.5×10^6) with externality values of $(\alpha\beta = 0.53)$. In other words, the cloud platform attracts (3.3×10^6) of data consumers by providing 4800 units of computing infrastructure at externality values of $(\alpha\beta = 0.2)$, while it attracts (0.5×10^6) data consumers by providing the same number of computing infrastructure units with externality values of $(\alpha\beta = 0.53)$. Consequently, in such a case, the cloud platform asks for a higher portion χ_i of revenues to maximize its payoff, which negatively affects the payoff of the data providers. In Figure 3.10, we study the impact of the externality parameters on the portion of revenues asked by the cloud platform. As shown in the figure, the cloud platform asks for higher portions as the values of externality parameters are increased. For example, at a subsidizing factor of $\phi = 1.5$ and externality values of $(\alpha\beta = 0.3)$, data providers share 40% of their revenues, while they share 68% of their

revenues with externalitiy values of $(\alpha\beta = 0.6)$ and a constant subsidizing factor of $\phi = 1.5$.

3.4.4 Sensitivity Analysis of Subsidizing Factor and Greedy Behavior of Involved Parties

The exponential function χ_i^{ϕ} captures the rational/greedy behavior of the cloud platform and data providers. The subsidizing factor ϕ implicitly describes the reactions of the cloud platform to the sharing portions offered by data providers. Theoretically, the cloud platform subsidizes the data providers by imposing a subsidizing factor ϕ that is less than 1. This leads to having $\chi_i^{\phi} > \chi_i$ since the base χ_i is defined to be between 0 and 1, meaning that larger amounts of computing infrastructure units D_{s_i} should be provided. On the other hand, data providers offer higher portions χ_i when the cloud platform acts greedily by imposing a subsidizing factor ϕ that is greater than 1. This leads to having $\chi_i^{\phi} < \chi_i$, meaning that smaller amounts of computing infrastructure units D_{s_i} need to be provided. When the subsidizing factor is equal to 1, the data providers offer a market share to the cloud platform based on the size of its contribution. This leads to having $\chi_i^{\phi} = \chi_i$, meaning that the provided computing infrastructure is linearly probational with respect to the market share χ_i. In this case, neither the cloud platform nor data providers act greedily, and they do not subsidize each other at the same time.

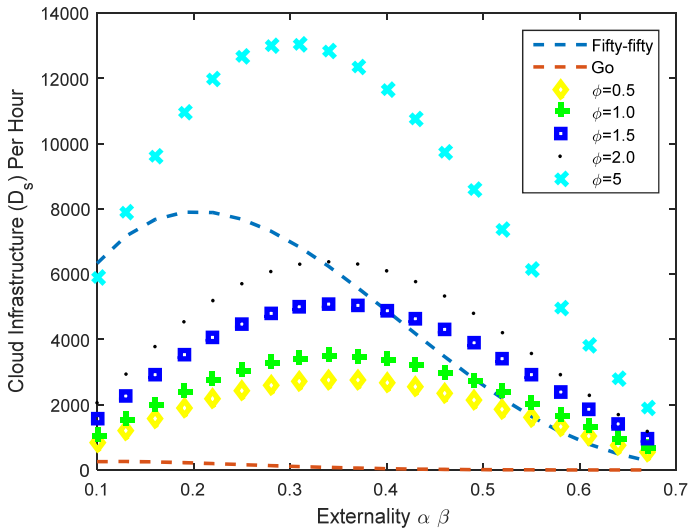


Figure 3.9: Cloud Infrastructure over externalities $\alpha\beta$

In Figures 3.6, 3.7 and 3.8, we consider the impact of the subsidizing factor ϕ in the sensitivity

analysis of the externalitiy parameters. As can be observed from those figures, the cloud platform and data consumers receive higher payoff than the data providers as the subsidizing factor increases. For example, in Figure 3.6, the cloud's payoff is higher under a subsidizing factor of $\phi = 2.0$ than it is under a subsidizing factor of $\phi = 1.0$. Similarly, we can observe in Figure 3.8 that data consumers' demand is higher under a subsidizing factor of $\phi = 2.0$ than it is under a a subsidizing factor of $\phi = 0.5$. Unlike the cases of cloud and data consumers, data providers' payoff is less under $\phi = 5.0$ than it is when $\phi = 1.5$, as depicted in Figure 3.7.

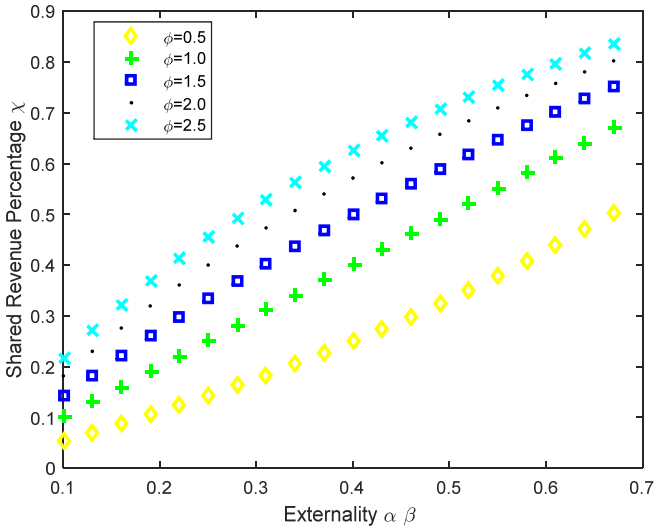


Figure 3.10: Shared revenue among the cloud and data providers (χ_i) over externalities $\alpha\beta$

To better clarify the impact of the subsidizing factor on the payoffs of involved parties in our model, we run the simulation over a continuous, reasonable and wider range of subsidizing factor values. Specifically, we describe the payoff of the cloud, data providers and data consumers as a function of the subsidizing factor ϕ in Figures 3.11, 3.12 and 3.13 respectively. The results shown in these figures confirm the insights extracted from Figures 3.6, 3.7 and 3.8. Furthermore, we notice in Figures 3.12 and 3.13 that the payoff of the data providers and consumers increases when the value of ϕ is between 0 to 1. Then slightly decreases /stabilizes as ϕ becomes greater than 1. On the other hand, the cloud's payoff, as shown in Figure 3.11, largely increases when ϕ becomes greater than 1.

This behavior can be practically interpreted with the results shown in Figure 3.10. In this figure, we notice that data providers react to the greedy behavior of the cloud platform via increasing the

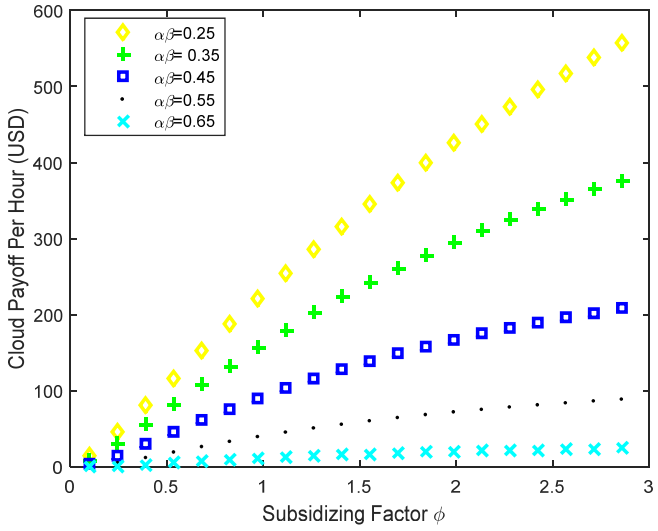


Figure 3.11: Cloud payoff over the subsidizing factor ϕ

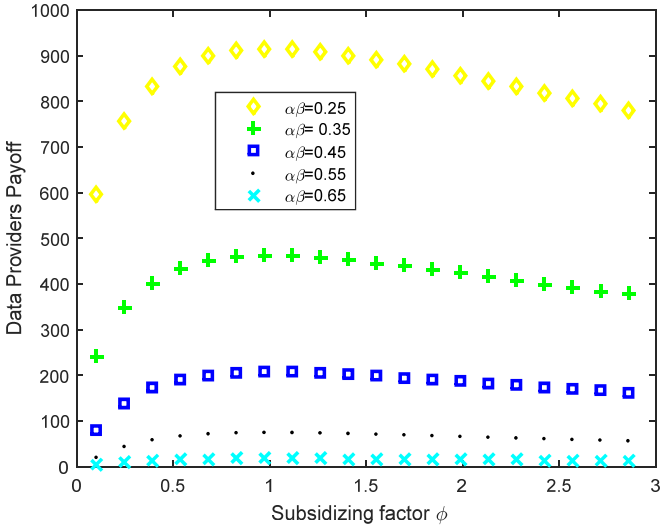


Figure 3.12: Data providers payoff over the subsidizing factor ϕ

shared revenue, which negatively affects the payoff of data providers as shown in Figures 3.7 and 3.12, but positively affects the cloud's payoff as shown in Figures 3.6 and 3.11.

However, the behavior of the data providers (reacting to the greedy behavior of the cloud platform via increasing the shared portions) sustains a higher level of consumers' demand as shown in Figures 3.11 and 3.8. Similarly, as shown in Figure 3.10, the cloud platform acts to the low offered portion by imposing a subsidizing factor ϕ that is less than 1. This negatively impacts the cloud's

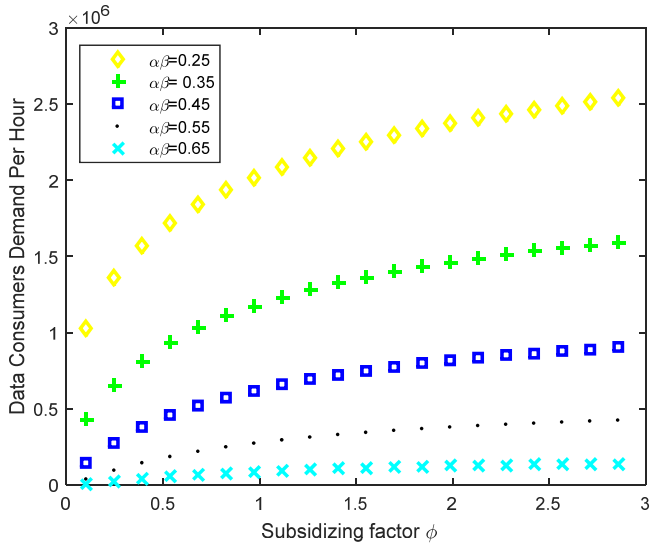


Figure 3.13: Consumer demand over the subsidizing factor ϕ

payoff but positively affects the cloud platforms' payoff. In summary, we conclude that our game includes a recovery mechanism that helps sustain efficient payoff outcomes for all the parties, in case any of the involved parties decides to act greedily toward the others.

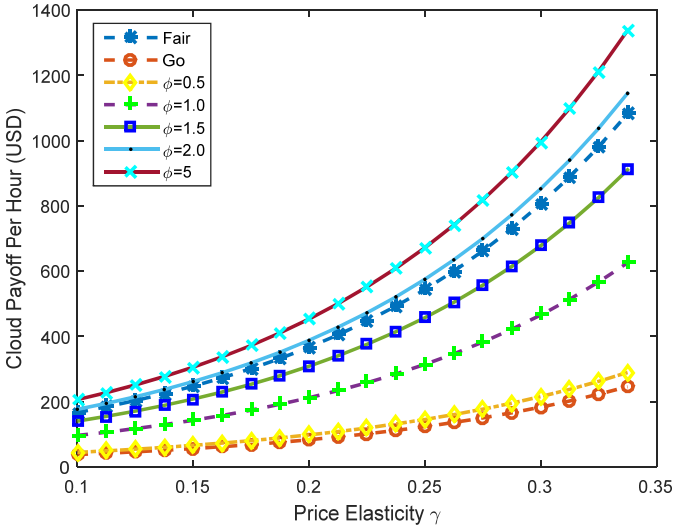


Figure 3.14: Cloud payoff over demand elasticity γ

3.4.5 Sensitivity Analysis of Consumer Demands Elasticity (γ) and the Multiplier (k_1)

We now move to analyzing the impact of consumers' demand elasticity on the surpluses of all involved parties (Figures 3.14 - 3.16). A high negative elasticity value positively impacts the surplus of all parties under the two-sided game model through yielding higher payoffs for the cloud platform and data providers. The reason is that the exponential function $P_i^{-\gamma}$ increases when γ increases from 0 to 0.34 as long as the base P_i is less than 1. This makes data providers able to charge their consumers higher prices without significantly leading to a decrease in the demands.

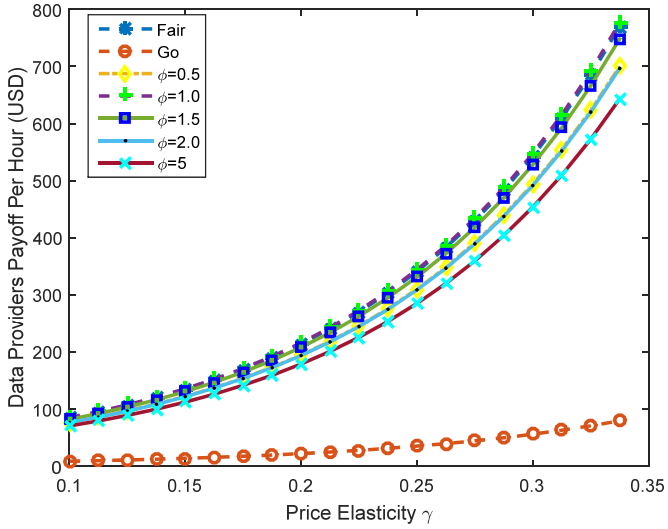


Figure 3.15: Data providers payoff over demand elasticity γ

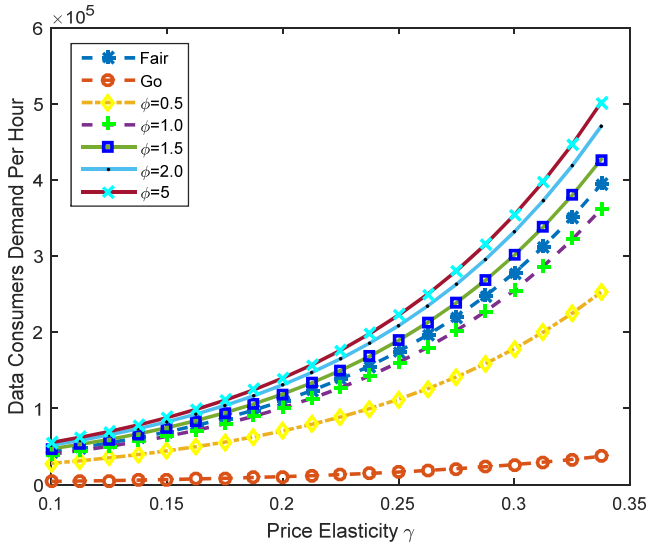


Figure 3.16: Consumers demand over demand elasticity γ

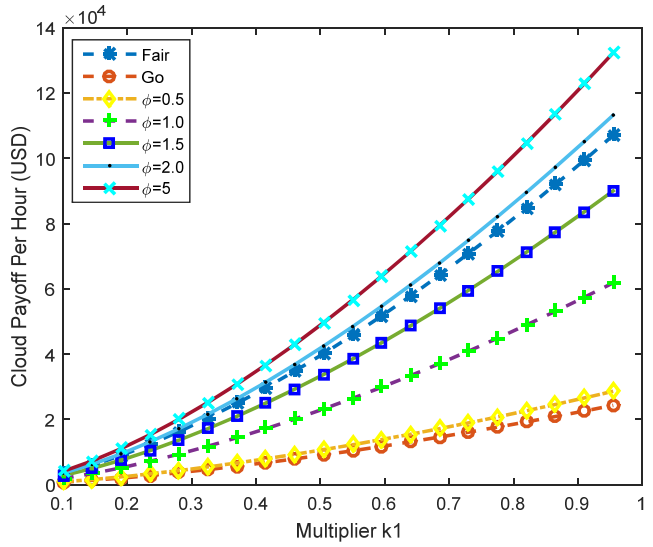


Figure 3.17: Cloud payoff over multiplier k_1

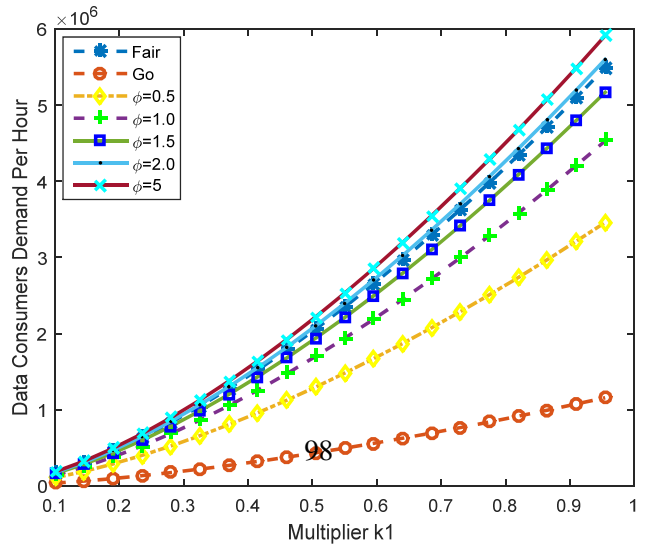


Figure 3.19: Consumers demand over multiplier k_1

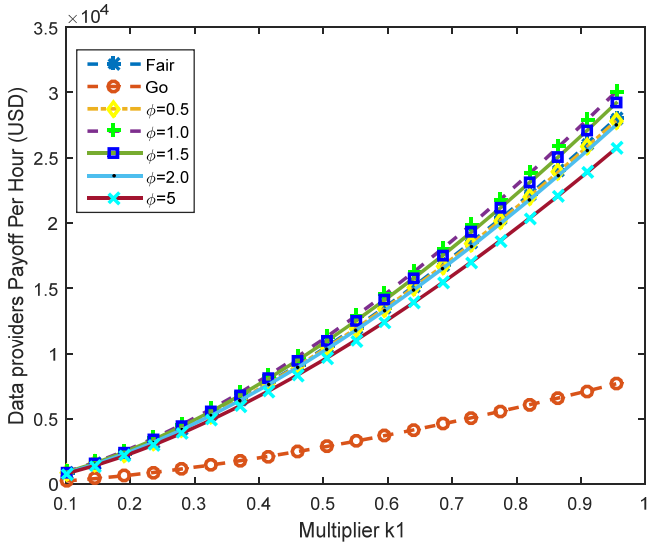


Figure 3.18: Data providers payoff over multiplier k_1

We also study the impact of the multiplier k_1 on the surplus of all involved parties. As shown in Figures 3.17 - 3.19, the surpluses of the cloud platform, service data providers and consumers increase as the value of the multiplier k_1 increases. The reason is that increasing the multiplier k_1 leads to increasing consumers' demands, which positively affects the total revenues.

3.5 Conclusion

In this paper, we proposed a game theoretical model based on the two-sided market theory to monetize big data services over the cloud. In particular, we studied the impact of cross group externalities between the amount of provided cloud resources and number of service consumers. The objective is to come up with a new vision in which the cloud can play a primordial role in introducing big data service providers and data consumers to each other, which results in higher benefits for all the involved players. To achieve this goal, we designed a game theoretical model in which the cloud platform and big data service providers engage in a closed loop of dependencies that makes them interested in satisfying each other interest, instead of following an aggressive competition strategy. Empirical results showed that our model outperforms the state-of-the-art cloud business models, i.e., the fifty-fifty and pay-as-you-go models in terms of total surpluses earned by the different parties.

Chapter 4

Cloud as Platform for Monetizing Complementary Data for AI-driven Services: A Two-Sided Cooperative Game

In this paper, we design a strategic game that aims to deliver complementary data services among multiple data providers over a cloud intermediary platform. More specifically, we formalize the problem as an extended two-sided market model by courting on one side some influential data providers in order to attract other data providers on the same side to form a bundling of data services. Simulation results using real data sets from Amazon and Google cloud show promising results for the different involved agents in terms of total surpluses and maximum revenues in different settings.

4.1 Introduction

Nowadays, Artificial Intelligence (AI)-driven services are being used in many industries and sectors such as driver-less cars, medical care, finance, etc. In general, AI-driven services release technology solutions to assist organizations and individuals by executing machine learning and data analytics procedures on massive data involving multiple data types, generated by multiple data providers. For example, Riskified, an AI-driven recommendation service helps e-commerce sites release new products and enter new markets as well as identify legitimate shoppers. Riskified

requires more than one billion past transactions including data about the products, stores, user's purchases, brands, and associated data about the customers to make excellent instant decisions. Such inherently combinatorial datasets which are formed by integrating different data types from multiple data providers are referred to in our paper as complementary data.

However, the research communities expect a turn down in the revolution of AI-driven services due to the shortage in the availability of big complementary data that need to be (pre-)trained using machine learning algorithms [106]. Specifically, AI-driven services entails high costs associated with collecting and integrating the big complementary data scattered across foundations and countries. Moreover, finding and getting on board multiple data providers raises management challenges associated with the level of collaboration, participation and consensus among the data providers to deliver a bundle of complementary data.

In this work, we argue that cloud computing can play a critical role in solving the challenge of providing complementary data for AI-driven services. The cloud hosts an explosive amount of data coming from a variety of enterprises and manufacturers that are deployed on its computing platforms. For example, the study reported in [2] revealed that one million customers deploy their own enterprises on Amazon, spending 30 billion USD on persistent storage on Amazon EC2 instances and generating 600 ZB of data per year [46]. Thus, the cloud computing could be used to liberate AI-driven services from having to search and discover appropriate data providers and to give them the opportunity to extract valuable patterns of information from massive complementary data, originating from multiple data providers.

This paper addresses the following challenges to achieve our purpose of enabling the cloud computing to monetize the complementary data for AI-driven services. **Challenge 1:** the cloud computing is not the actual owner of the data and has no right to monetize them directly without considering their actual owners. On the other hand, the idea that the cloud purchases the data from the data providers, and thereafter releases it for AI-driven services in exchange of a certain payment, which is known by the merchant model, is strongly criticised by law and research communities [16] for two reasons; 1) laws across the world such as the General Data Protection Regulation (GDPR) imposes high restrictions in terms of data privacy and security on selling the users' data for a third party. Alternatively, laws are strengthening regulations to move the control of data selling to the

hands of its actual owners; 2) the merchant model entails an aggressive competition between data providers and the cloud computing concerning the data price, which leads to fail in the agreement among involved parties or coarse distribution for surpluses. **Challenge 2:** the data types in complementary datasets exhibit a range of correlations and dependencies in the sense that the availability of a certain data type impacts the monetary value of other data types. This challenge raises the need to design a pricing scheme that estimates the value of a certain data type in the presence of other data types participating in the same complementary dataset. **Challenge 3:** irrationality, preferences and conflicts of interests of data providers raise the need to design a strategic mechanism accommodating and addressing the potentially undesired data provider's choices in terms of price and size of the data provided for the training.

To address the aforementioned challenges, we propose an innovative business model that views the cloud computing as an active two-sided market platform which gets on board complementary data providers and AI-driven services. The business model is supported by a bundling game-based strategy for complementary data that tackles **challenge 2** and **challenge 3**. In this game, the cloud computing acts as an orchestration platform that considers and regulates the strategies of the data providers to increase the profit of all involved parties (i.e., cloud computing, data providers and AI-driven services). The details of the research problem and contributions are explained in Section 4.2. Section 4.3 introduces our model for bundling complementary data. The section discusses as well the model formulation and utility functions. Section 4.4 presents the simulation setup and discusses the simulation results. Section 4.5 discusses the related work. The paper is concluded in Section 4.6.

4.2 Motivating Scenario and Technical Contributions

4.2.1 Cloud as a Two-Sided Market

As mentioned earlier, the cloud computing is not the actual owner of the data and has no right to monetize them directly without considering their actual owners. Alternatively, the cloud computing can act as active global two-sided market which gets on board the AI-driven services as data consumers with data providers (actual owners) to execute machine learning and data analytics

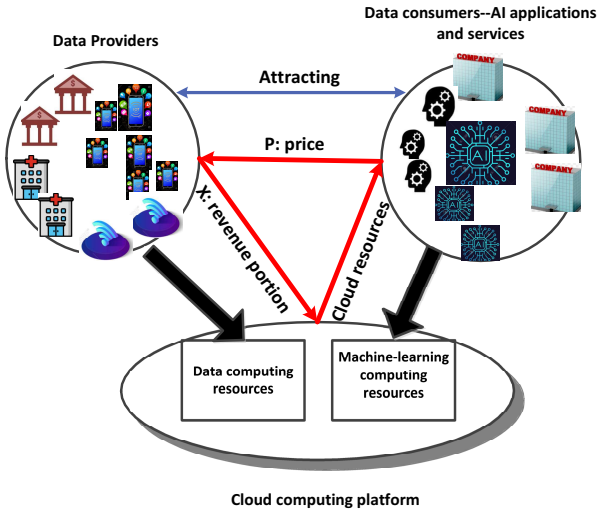


Figure 4.1: Cloud as platform for data and AI services

processes as depicted in Figure 4.1. The two-sided cloud computing platform for monetizing complementary data for AI-driven services plays two key roles: 1) introducing wide social networks of AI-driven services to the data providers and vice versa; and 2) providing computing infrastructure for both the AI-driven services to deploy their machine learning and data analytics procedures and data providers to deploy their collected data. Consequently, AI-driven services will be attracted to the cloud platform to which a massive number of data providers are connected and vise versa. The AI-driven services pay the data providers versus executing their machine learning procedures on the providers' data; while data providers share a monetary reward or a portion of their revenues with the cloud computing platform. The objective of the cloud platform is to maximize its profit by maximizing the transactions between the AI-driven services and data providers while considering the associated costs. Under the settings of this scenario, the cloud computing platform can choose from the two following strategies (depicted in Figure 4.2): 1) attracting the data providers by asking a low portion of their revenue, which in turn attracts the AI-driven services as followers; or 2) attracting the AI-driven services by providing them with highly performing computing resources, which in turn attracts the data providers as followers.

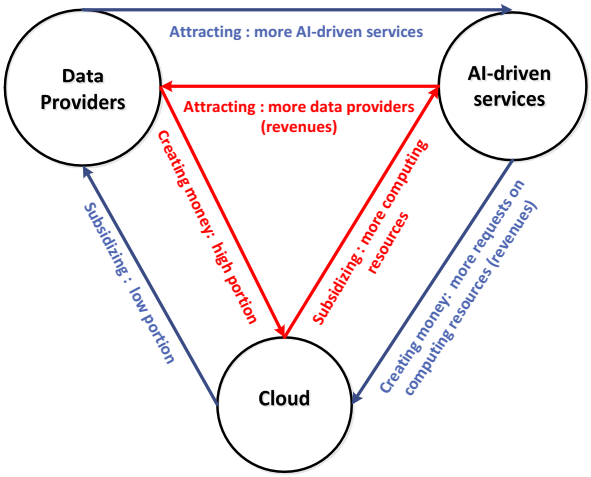


Figure 4.2: Two-sided game: Cloud strategies

4.2.2 Motivating Scenario

Suppose a scenario where a software company is seeking to develop an AI-driven service that assists pharmaceutical companies to develop treatments (e.g., drugs, vaccines, etc.) for the people that are diagnosed with COVID-19. To do so, the developers require relevant complementary datasets containing data from hospitals as well as some data about the environment (e.g., pollution rates, temperature rates, etc.) and data from wearable devices that monitor the health conditions of their users. In this scenario, the data from hospitals are the most influential among the other types of data. Moreover, hospital data are required to extract a certain patterns of information from the data about the environments and the ones coming from wearable devices. An example of this might be the AI-driven service investigating the impacts of the environment on the active periods of COVID-19, and the impact of the characteristics of the immune system observed by wearable devices on the likelihood of being infected with COVID-19. Consequently, the AI-driven service would have less demand on environmental and wearable device data in presence of insufficient amounts of hospital data. Therefore, the AI-driven service will be willing to pay more and to connect to the cloud to find data from hospitals. This highlights the externalities among the data providers and the AI-driven service, and justifies the need for the cloud as a two-sided market platform [82]. However, patients'

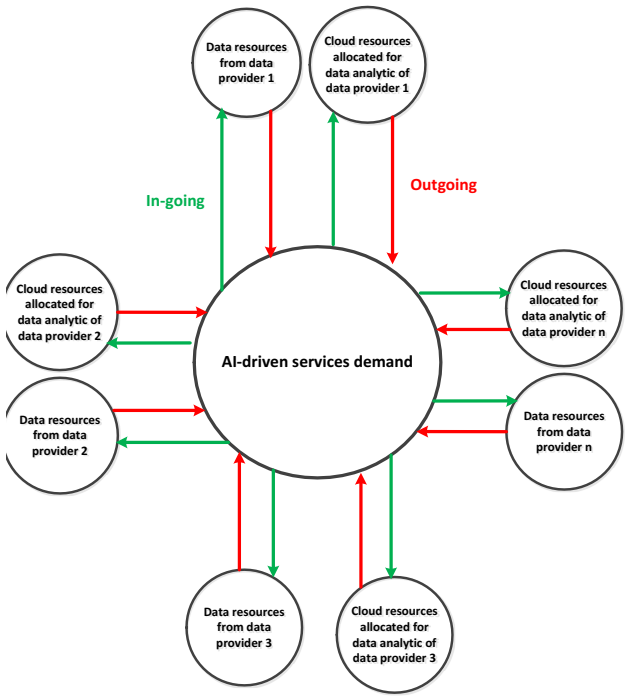


Figure 4.3: Dependencies among providers

data in hospitals are very sensitive and entail high operational cost in terms of machine learning computing resources. This leads to a lower supply from the hospitals, and hence a lower demand on the environmental and wearable device data. To address this challenge, we allow in our solution the cloud platform to subsidize the hospitals by offering low or even free computational fees for their machine learning computing resources to sustain an efficient supply of patients' data. This, in its turn, leads to an increase in the AI-driven services' demand and raises their willingness to pay more to acquire the data. In addition, this subsidizing strategy raises the willingness of environmental and wearable device data providers to connect to the cloud and pay more versus deploying their data enterprise on this cloud platform. This raises the need for a bundling data mechanism that selects the data providers that need to participate in the bundled data services and determines the optimal amount of computing resources for each one. Economically speaking, the absence of such a mechanism might lead to less total surpluses for all the involved players.

4.2.3 Technical Contributions

In view of the motivating scenario, the cloud computing needs to answer the following questions: 1) Which data providers will be selected to participate in the bundling strategies? 2) Which ones of the selected providers will be subsidized? 3) How the selected data providers will be incentivized to sustain a maximum revenue and prevent undesired and greedy behaviour? Similarly, from the data providers' perspective, the following questions need to be answered: In the presence of other data providers participating in the bundled complementary data, 1) How should a provider price its own data service? 2) How much should the provider pay to the cloud computing versus the computational fees?

To help the cloud platform and data providers answer these questions, we propose in this work a game theoretical model based on the two-sided market theory. The cloud intermediates the interactions between the AI-driven services from one side and data providers from the other side. Our model allows the cloud platform to subsidize the most influential data providers with (1) a low portion of revenues and (2) machine learning computing resources to offer them an efficient level of supply. By getting on board the most influential data providers over the cloud, the other data providers, whose data become more valuable when combined with those of the influential subsidized ones, will be attracted to perform their data transactions over the same cloud platform. They will also be more encouraged to pay higher fees to the cloud platform in order to join forces with the influential providers. The resulting bundled data services appeal to a higher number of AI-driven services, thus leading to higher levels of revenues to both data providers and the cloud platform.

In the above-described architecture, the cloud acts as an orchestration player that controls the supply of data services into the bundled data services using the allocated machine learning computing resources for each service, as depicted in Figure 4.3. For example, the cloud may exclude a greedy data provider that demands a high price for its service from the bundled data services by not allocating the needed machine learning computing resource for its computational data analytics.

The dependencies among the data providers (i.e., the effects of each data provider's actions on the rest of the providers) are modeled indirectly as cross-group externalities between AI-driven services and data providers. As shown in Figure 4.3, a data provider affects the demand on the

AI-driven services by an outgoing externality, and consequently, all the other data providers will get affected by an in-going externality. The outgoing externality refers to the raise in the AI-driven service's demand when one more additional data unit is supplied by the data provider. The in-going externality refers to the impact of having an additional AI-driven service on the payoff of the data provider. On the other side, the externalities between the cloud platform and AI-driven services are taken into account to model the orchestration role of the cloud. Specifically, the cloud distributes its machine learning computing resources over the data providers. In this case, the outgoing externalities refer to the raise in the AI-driven service's demand when one more additional machine learning computing unit is allocated by the cloud to speed up the process of executing a certain data analytics procedure. The in-going externalities refer to the impact of having one additional AI-driven service on the net payoff of the cloud platform that results from executing a certain data analytics procedure.

At the technical level, we contribute to the two-sided market theory in the following respects: 1) We consider the dependencies among the players on one side (i.e., data providers' side); and 2) we add a two-stage subsidizing: one to attract some data providers, and another one to attract the AI-driven services. On top of this, we embed double cross-group externalities across the players as follows: a) cross-group externalities between data providers and AI-driven services, which are common in two-sided markets; and b) cross-group externalities between the cloud platform and the AI-driven services. The latest one adjusts the power balance among the players (i.e., cloud and data providers) and alleviates the competition between them. In these proposed settings, we derive the new equilibrium of our two-sided market scenario. To the best of our knowledge, our work is the first that capitalizes on the two-sided market theory as a platform to get on board services for complementary data.

4.3 Proposed Model for Bundling Complementary Data Services

4.3.1 Model Entities Description

The proposed data bundling model, depicted in Figure 4.4, is composed of three main entities: data providers (SP), AI-driven services (AIS) and the Cloud Computing platform (CC). The cloud computing platform consists of a *master cloud computing node*, *edge computing servers*, and *base*

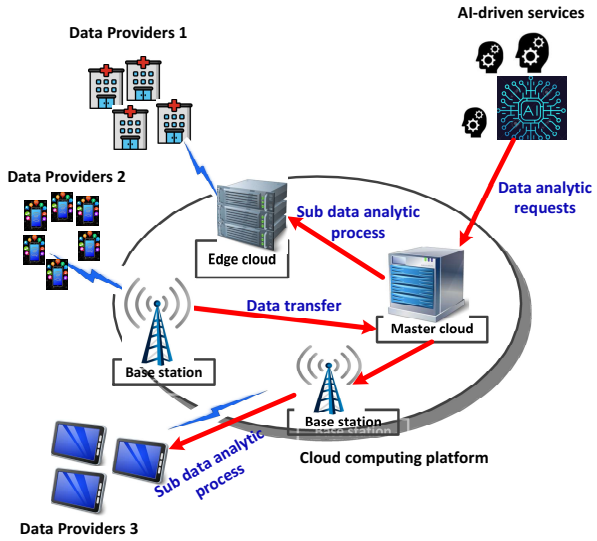


Figure 4.4: Data bundling scenario over cloud computing

station servers. The requests made by AI-driven services for a bundled data service are handled using multiple data analytics sub-processes (SC) that get established among the underlying AI-driven services and data providers over the cloud computing platform. Each sub-process among a data provider, an AI-driven service and the cloud computing includes negotiations about the quality of the provided services (in terms of data transaction throughput), data type, payments, as well as all the other terms of the delivered data services. The AI-driven services' demand on a bundle of data services supplied by multiple data providers is denoted by Dc.

The data providers vary in the professionalism level of data providing. For example, organizations, such as hospitals in Figure 4.4, mainly join the cloud to rent some computing infrastructure to deploy and support their IT systems. The objective is to assist their staff with the creation and maintenance of the software and configuration associated with their internal business with minimal effort. Active daily transactions and interactions over those IT systems generate beneficial data associated with the businesses of those organizations. Consequently, the generated data attracts AI-driven services and the role of those organizations starts to become data providers instead of being cloud customer solely. The data, in such a case, already resides on the cloud computing servers. Consequently, the machine learning procedures of the AI-driven services are executed over the cloud computing servers, which aligns with the centralized machine learning style.

Another example of data providers are professional data providers whose businesses mainly

concern collocating and processing the data. For example, the IoT service providers collect and process sensing data in order to provide associated services in different domains. This type of data providers may have their IT equipments such as computing servers, and the data resides over their own computing resources. Such data providers, once being involved in the proposed model of monetizing complementary data for AI-driven services, mainly connect the cloud computing to get benefit from the wide social network of AI-driven services, while the machine learning procuders of AI-driven services are executed over its own computing infrastructure. Consequently, there is no need to transfer the (pre-) trained data to cloud servers, which aligns the federated learning style.

To consider this variety of professionalism level of data providers, the cloud computing platform can execute two machine learning approaches: 1) **Federated- learning data analytics:** no need to transfer the data for cloud servers. This style is presented in Figure 4.4 by the sub data analytics process between data providers 3 and AI-driven services over the master cloud computing node. A data provider SP_i contributes to each data service i by providing both data instances D_i and (machine learning-dedicated) computing resources Ds_i, which are used by AI-driven services to execute the data analytics tasks. The computing resources Ds_i are measured in terms of generated computing cycles per second. The role of the cloud computing platform in this scenario, beside introducing the social network of AI-driven services, is to provide an efficient level of communication bandwidth Dt_i between the data provider SP_i and cloud platform CC while exchanging the learning models. For example, in Figure 4.4, the base station which belongs to the cloud computing reserves a communication bandwidth between the data providers 3 and the master cloud to exchange the learning model. The communication bandwidth Dt_i is measured by the size of transferred data per second ($MB/second$). The average available energy E_i for data provider SP_i is a critical factor in this scenario, especially if the computing nodes are IoT devices with limited resources. 2) **Centralized-data analytics:** the data already resides on the cloud servers or there is a need to transfer it for the cloud server. This style is presented in Figure 4.4 by the two data analytics sub-processes between data providers 2, data providers 1 from one side and data providers 2 and AI-driven services from the other side over the master cloud computing. A data provider SP_i contributes to each data service i in this machine learning style by providing only the data instances D_i, which are transferred to the *master cloud computing node*. For example, data providers 1 deploy their data on the edge

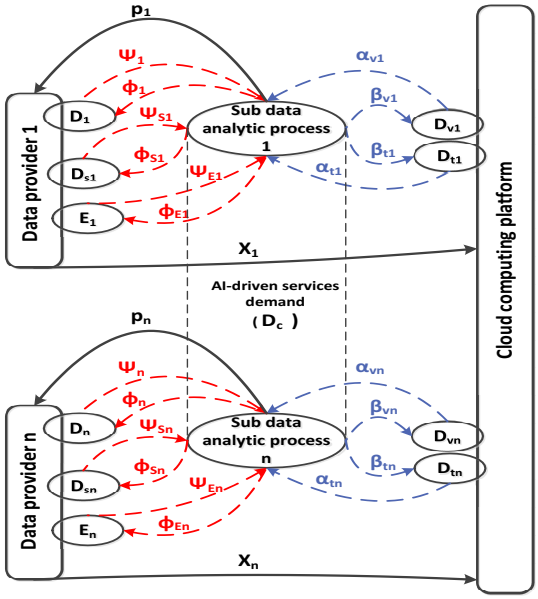


Figure 4.5: Two-sided cooperative model

server as shown in Figure 4.4. The role of the cloud computing in this scenario is mainly to provide computing resources Dv_i such as virtual machines for running machine learning-related tasks. The computing resources Dv_i are measured in terms of generated computing cycles per second.

4.3.2 Model Formulation: An Orchestration Two-Sided Cooperative Game

To model the relationships among the service providers SP, AI-driven services, and cloud computing platform, we employ the two-sided market theory [82] as depicted in Figure 4.5. The AI-driven services' demand (Dc) on a certain bundled service of complementary data is affected by all the data analytics sub-processes associated with the data providers. In its turn, a data analytics sub-process i associated by the data provider i is affected by the involved resources of the cloud computing platform and data provider i. The relationships among data provider SP_i and AI-services' demand Dc is implemented by **Data providers - AI-driven services cross group externalities**. These types of externailities are denoted in the model by the symbols ψ and ϕ. ψ refers to the impact of a data provider on the AI-driven services demand while ϕ denotes the impact of the AI-driven services' demand on the data provider. The symbols ψ and ϕ are presented by the out-going externality and in-going externality respectively as depicted in Figure 4.3. The

data provider SP_i may provide three types of resources: 1) **Data instances (D_i)** with ψ_i being the increase in the number of AI-driven services' requests when additional data units are invested and ϕ_i representing the increase in the amount of data units (D_i) per each additional request from AI-driven services. 2) **Machine learning computing resources (Ds_i)** with ψ_{si} being the increase in the number of AI-driven service requests when additional computing resources are invested and ϕ_{si} being the increase in the amount of computing resources (Ds_i) per each new additional AI-driven service request. The externality ψ_{si} captures in our model the computational overhead and the complexity of the machine learning procedures of the AI-driven services executed over the data instance D_i. For example, the data provider SP_i reserves more computing resources Ds_i to handle machine learning tasks faster. The relatively high value of the externality ψ_{si} refers to the importance of the required time to accomplish a certain machine learning procedure. The externality ϕ_{si} in our model captures the economical impact in terms of the revenues and costs, associated with increasing the AI-driven services requests on the supply of computing resources Ds_i. The relatively high value of ϕ_{si} refers the high beneficial economic revenue and the high willingness of the data provider SP_i to assign more computing resources Ds_i. The positive value of the externalities ψ_{si} and ϕ_{si} indicates that the data analytics sub-process SC_i aligns with the federated learning style, where the machine leaning procedures of the AI-driven services are executed on the computing resources of the data provider SP_i without the need to move the data to the cloud servers. The zero value of the externalities ψ_{si} and ϕ_{si} refers for the centralized learning style where the data provider SP_i engages in the monetization process by only the data instance Ds_i that is either already deployed or needed to be moved for the cloud servers. 3) **Energy resources (E_i)** with ψ_{Ei} representing the increase in the number of AI-driven services' requests when additional energy units are invested and ϕ_{Ei} representing the increase in the energy resources (E_i) per each new additional AI-driven service request. The relatively high value of the externality ψ_{Ei} refers to the high computational overhead and complexity of the machine learning procedures that require higher energy levels. The relatively high value of ϕ_{Ei} refers for the highly profitable economic revenue and the high willingness of the data provider SP_i to offer more energy E_i. Similar to the externalities ϕ_{si} and ψ_{si}, the positive value of the externalities ψ_{Ei} and ϕ_{Ei} refers for the federated learning style, while the zero value refers for the centralized learning style.

Having a cross-group externality representation of AI-driven services' demand with regard to data providers helps us also to seize the impacts of each data provider on the rest of the providers. Technically speaking, the data provider SP_i will have to maximize the demand of the AI-driven services on the offered bundled services in order to maximize its own profits. Increasing this demand (from one provider's side) indirectly results in increasing the revenues of the other providers, which boosts a cooperative behavior in the bundled data service.

Similarly, the interactions among the demand of the AI-driven services and the supply of the cloud computing is modeled as **Cloud computing platform - AI-driven services cross group externalities**. These types of externailities are denoted in the model by the symbols α and β. α refers to the impact of a cloud computing platform on the AI-driven services demand while β refers to the impact of the AI-driven services demand on the cloud computing platform. α and β are presented by the out-going externality and in-going externality respectively in Figure 4.3. The cloud computing platform contributes to the sub data analytics process i by two types of resources: 1) **Virtual machine learning computing resources** (Dv_i) and 2) **Bandwidth communication resources** Dt_i. α_{vi} and α_{ti} quantify the increase in the number of AI-driven services when some additional computational resources and bandwidth power are added to Dv_i and Dt_i respectively. β_{vi} and β_{ti} quantify the amount of benefits that the cloud computing platform yields when new AI-driven services' request are added to Dc. The relatively high value of the externality α_{vi} refers to the highly centralized computational overhead of the machine learning procedures that are executed over the cloud servers. The zero value of the externalities α_{vi} and β_{vi} refers for the federated learning style where the data and machine learning procedures are deployed and executed respectively at the provider's side and not at the cloud's side. Similarly, The relatively high value of the externality α_{ti} refers to the high communication overhead, which is usually entailed by the federated learning style in terms of exchanging the learning models.

The cross-group externalities seize the capabilities of the cloud computing as an orchestration player in terms of ability to 1) manage and control the cooperative data providers participating in the bundled data service, and 2) sustain the maximum payoff of the players. Specifically, the cloud computing uses the allocated computational resources Dv_i and Dt_i to control and govern the throughput of the data transactions contributed to the bundled data service by each data provider

SP_i. For example, let's assume that a greedy data provider SP_i sets a high price P_i for its contributions in a data bundled service. This negatively affects the overall AI-driven services' demand on the bundled service. Consequently, the cloud computing would stop or at least slows down the data analytics SP_i by assigning it a lower portion of machine learning computing resources Dv_i. As a result of the cloud's reaction, the data provider SP_i would be either excluded from the bundled service or it would receive a smaller portion of the revenues. Thus, the strategy of the cloud aims to force the data provider SP_i to lower its price and behave in a non-greedy fashion to sustain a maximum level of revenue and stay in the bundled service.

On the other hand, the interaction between service providers SP and cloud computing CC is modeled as a two-stage game. In this game, CC takes the role of game leader and SP act as its followers. In the first stage of the game, each data provider SP_i responsible of providing service i observes the data providers participating in the bundled data service and the amount of money returns X_i requested by CC. This allows the provider to tune the supply volume of computing resources and the price that will be charged to the AI-driven services of the same service i. Based on the price specified by SP_i, CC defines the optimal amount of computing resources D_{s_i} to be offered to regulate the data analytics between SP_i and AI-driven services. The proposed model takes the form of a closed loop of dependencies and includes a subsidizing technique from the two-sided market theory. Specifically, in our game, CC may subsidize SP_i with a low fraction of its revenue to maintain an optimal level of P_i or to entice other data providers.

4.3.3 Players Demands and Utility Functions

To model the AI-driven services' demand and supply, we employ the Cobb-Douglas function. The advantage of this function is that it can capture the elasticity of the computing/storage resources supply (D_{s_i}, D_i, E_i, Dv_i, and Dt_i) with respect to the user's demand. The demand functions are given using Equations (51), (52), (53), (54), (55), and (56), where: 1) k_i, Ks_i, Kc, Kv_i, and Kt_i are constant multipliers; 2) γ_i is Dc's elasticity with respect to P_i; 3) ω_i is SP_i's elasticity with respect to shared revenue X_i.

$$Dc = Kc \prod_{j=1}^{n} P_j^{-\gamma_j} D_j^{\psi_j} Ds_j^{\psi_{sj}} E_j^{\psi_{Ej}} Dv_j^{\alpha_{vj}} Dt_j^{\alpha_{tj}} \tag{51}$$

$$Ds_i = Ks_i X_i^{-\omega_i} (P_i Dc)^{\phi_{si}} \tag{52}$$

$$D_i = k_i X_i^{-\omega_i} (P_i Dc)^{\phi_i} \tag{53}$$

$$E_i = k_{Ei} X_i^{-\omega_i} (P_i Dc)^{\phi_{Ei}} \tag{54}$$

$$Dv_i = Kv_i (X_i P_i Dc)^{\beta_{vi}} \tag{55}$$

$$Dt_i = Kt_i (X_i P_i Dc)^{\beta_{ti}} \tag{56}$$

By mapping Equations (52), (53), (54), (55), and (56) into Equation (51), we can represent the AI-driven services' demand as a function of X_i and P_i as shown in Equation (57). where: 1) a_{1i} $= \alpha_{vi}\beta_{vi} + \alpha_{ti}\beta_{ti} - (\psi_i + \psi_{si} + \psi_{Ei})\omega_i$; 2) $a_{3i} = \psi_i\beta_i + \psi_{si}\beta_{si} + \psi_{Ei}\beta_{Ei} + \alpha_{vi}\beta_{vi} + \alpha_{ti}\beta_{ti}$; 3) $a_{2i} = a_{3i} - \gamma_i$; and 4) $a_4 = 1/(1 - \sum_{j=1}^{n} a_{3j})$

$$Dc = (Kc \prod_{j=1}^{n} k_j X_j^{a_{1j}} P_j^{a_{2j}})^{a_4} \tag{57}$$

Each data provider SP_i should pay a constant cost Fs_i per a single AI-driven service's access. That being said, SP_i attempts to maximize its payoff that is depicted in Equation (58).

$$\pi_i = ((P_i)(1 - X_i) - Fs_i)Dc \tag{58}$$

The cloud computing CC should pay a constant cost Fc_i for each sub process SC_i it establishes between a service provider and an AI-driven service. As a result of being rational, the cloud

computing attempts to maximize its payoff which is described in Equation (59).

$$\pi = (\sum_{j=0}^{n}(X_j P_j - Fs_j))Dc \tag{59}$$

4.3.4 Game Equilibrium

We derive the equilibrium of our game using backward induction. In fact, the sub-game of the service providers (followers of the game) is initially solved to derive their optimal response P_i^* to the AI-driven services. The sub-game of the cloud computing (leader of the game) is then derived to get the optimal X_i^*. Using the solutions of these two sub-games, the equilibrium of the game is expressed in Theorem 3.

Theorem 3. *The equilibrium of our orchestration two-sided cooperative game is derived from the best responses of the different players using the following methodology:*

(1) The data provider SP_i's best response is computed as follows:

$$P_i^* = \frac{a_4 a_{2i} Fs_i}{(1 + a_4 a_{2i})(1 - X_i^*)} \tag{60}$$

$$\textit{if:} \quad (1/(a_4 a_{2i})) > -1$$

(2) The cloud computing best response concerning service i is computed as follows:

$$X_i^* = \frac{-a_4 a_{1i} \sum_{j=1}^{n}(X_j p_j - Fc_j)}{P_i^*} \tag{61}$$

Proof. By applying the $\log$ on both sides of the data provider's payoff equation (Equation (58)), we obtain:

$$\log \pi_i = \log(P_i(1 - X_i) - Fs_i) + \log Dc \tag{62}$$

Then, the optimal price P_i^* is defined by $\partial \pi_i / \partial P_i = 0$ as follows:

$$\frac{1}{\pi_i} \times \frac{\partial \pi_i}{\partial P_i} = \frac{1 - X_i}{P_i(1 - X_i) - Fs_i} + \frac{1}{Dc} \times \frac{\partial Dc}{\partial P_i} = 0 \tag{63}$$

By deriving Equation (57) with respect to P_i, then:

$$\frac{\partial Dc}{\partial P_i} = a_4 a_{2i} Dc P_i^{-1} \tag{64}$$

By substituting Equation (64) into Equation (63), we get:

$$P_i^* = \frac{a_4 a_{2i} F s_i}{(1 + a_4 a_{2i})(1 - X_i^*)} \tag{65}$$

Since $(1 - X_i^*)$, Fs_i are positives, then $\frac{a_4 a_{2i}}{(1+a_4 a_{2i})} > 0$
$\Rightarrow (1/(a_4 a_{2i})) > -1$

Applying a first derivative test leads us to $\partial \pi_i / \partial P_i > 0$ when $P_i < P_i^*$ and to $\partial \pi_i / \partial P_i < 0$ when $P_i > P_i^*$ if $(1/(a_4 a_{2i})) > -1$. Consequently, we conclude that P_i is the best response.

For the second part of the theorem, we apply the $\log$ on both sides of the equation of the cloud's payoff (Equation ((59))) and obtain the following:

$$\log \pi = \log(\sum_{j=0}^{n}(X_j P_j - F c_j)) + \log Dc \tag{66}$$

Then, the optimal X_i^* is defined by $\partial \pi / \partial X_i = 0$ as follows:

$$\frac{1}{\pi} \times \frac{\partial \pi}{\partial X_i} = \frac{P_i}{\sum_{j=0}^{n}(X_j P_j - F s_j)} + \frac{1}{Dc} \times \frac{\partial Dc}{\partial X_i} = 0 \tag{67}$$

By deriving Equation (57) with respect to X_i, we get:

$$\frac{\partial Dc}{\partial X_i} = (a_4 a_{1i}) Dc X_i^{-1} \tag{68}$$

By substituting Equation (68) into Equation (67), then:

$$X_i^* = \frac{-a_4 a_{1i} \sum_{j=1}^{n}(X_j p_j - F c_j)}{P_i^*} \tag{69}$$

Since X_i^* and $\frac{\sum_{j=1}^{n}(X_j p_j - F c_j)}{P_i^*}$ are positives, then $-a_4 a_{1i} > 0 \Rightarrow a_4 a_{1i} < 0$

Applying a first derivative test leads us to $\partial \pi / \partial X_i > 0$ when $X_i < X_i^*$ and to $\partial \pi / \partial X_i < 0$ when $X_i > X_i^*$. Consequently, we conclude that X_i is the best response, which proves our theorem.

□

4.4 Simulation and Empirical Analysis

4.4.1 Simulation Objectives

In this experiment, we focus on demonstrating the efficiency of the proposed model in terms of the following questions: 1) How does the strength of dependencies (externailities) among data providers affect the shared revenue between both the data providers and cloud computing? 2) How does the cloud computing subsidize the data providers SP_i in terms of shared revenues over different ranges of dependencies ? 3) In the centralized learning style, how does the cloud computing subsidize the AI-driven services in terms of computing resources assigned for their machine learning procedures over different ranges of dependencies (externalities) ? 4) In the federated learning scenario, how does the strength of dependencies (externailities) among data providers affect the supply of machine learning computing resources assigned by SP_i for AI-driven services? Those questions have been answered in Section 4.4.3. 5) How does the strength of dependencies (externailities) among data providers affect the payoff of both the data providers and cloud computing? this question has been answered in Section 4.4.4. Some proposals in the literature, such as [16], executed empirical comparisons between the two-sided market and its bench mark, i.e., the merchant model. In this paper, we focus on other results.

4.4.2 Simulation Setup

We carry out our simulations according to the scenarios explained in Section 4.3.1, i.e., data providers engaged with the cloud to execute centralized or federated machine learning tasks, requested by a group of AI-driven services. In our simulations, the data provider SP_i chooses the price P_i of its data services from the interval $[0.2, 3.2]$ USD/hour, inspired by the distribution of prices for 150 IoT providers providing data services on Amazon's marketplace [9, 2]. We also capitalize on a dataset from Google [37], which contains statistics on the implementation of big data analytics tasks on Google-powered Virtual Machines (VMs). These statistics state that each VM needs an average of 1.42 to 10 seconds to fulfill a data analytics task (with a mean of 5.71s and a standard deviation of 4.29s). We denote the instances along with their average computational power in our model by Ds_i and Dv_i. Consequently, provisioning a compute instance, which is denoted in

our model by the externalities α_{si} and α_{vi}, leads to an increase between 0.1 to 0.7 in the number of data requests per second. As argued in [16], the cross-group externalities should be neither quite weak nor quite strong. The analysis in [16] revealed that the cross-group externalities should be in the range $0.1 < \alpha\beta < 0.6$. Consequently, the externality factor β would be bounded by $0.1/\alpha$ and $0.6/\alpha$. We assign the cross-group externalities ϕ and ψ with values from the same range of α and β. The elasticity γ of the price is fixed to 0.15, a value that is analogous to the sensitivity level of the price of mobile/telecommunication services as assessed in [31]. These above-described parameters are taken as inputs to our model, which then computes the optimal shared revenue X_i for each data i based on Equation (61) of Theorem 3 and other associated outputs such as cloud and data providers' payoffs.

4.4.3 Subsidizing Sensitivity

We investigate in this section the influence of the cross group externality metrics ($\phi_i\psi_i$). More specifically, we study the impact of the data on AI-driven services' demand on the shared generated revenue X_i for the data provider SP_i with regard to the cross-group externality metrics average ($\phi_{-i}\psi_{-i}$) of other data providers participating in the bundled service (the notation $-i$ refers to the other data providers different from SP_i). As shown in Figure 4.6, the cloud computing subsidizes the data provider SP_i by charging a lower percentage of the revenue as the externality factors $\phi_i\psi_i$ are relatively larger than $\phi_{-i}\psi_{-i}$ (i.e., the data provider SP_i is an influencer data provider while the others are followers). For example, the cloud computing charges around 35% of the shared revenue when the externality $\phi_i\psi_i$ is 0.6 and the average externality of other players $\phi_{-i}\psi_{-i}$ is 0.1. On the other hand, the cloud computing charges a higher portion (65%) of the shared revenue when the externality $\phi_i\psi_i$ is 0.6 and the average externality $\phi_{-i}\psi_{-i}$ is 0.6. However, the data provider SP_i pays a higher portion of shared revenue to the cloud computing provided that its externalities ($\phi_i\psi_i$) are low and that the other parties' externalities are high, (i.e., the data provider SP_i is a follower while others are influencers). This is due to the data provider's willingness to pay more to push the most influential data providers to join the bundled service. Nonetheless, the data provider SP_i pays a higher percentage of the shared revenue to the cloud as its externalities $\phi_i\psi_i$ become stronger.

Figure 4.7 captures the subsiding of the cloud for AI-driven sercvices in the centralized learning

style by assigning more computational units when the providers' data becomes less attractive for the AI-driven services. Similar observations can be drawn from Figure 4.8 with respect to the data providers. Precisely, the data provider SP_i assigns relatively more data computational units in the federated learning scenario, to attract more AI-driven services if its data type has less impact on the AI-driven services' demands. In addition, Figure 4.8 shows the effect of the data providers on each other in bundled services. As illustrated in the figure, the data provider SP_i invests more computational data units when the externalities of other data providers become stronger to stay in the strong bundled data service. This behaviour refers to an increase in the AI-driven services' demand on the provider SP_i's data when other data providers have strong externalities with the AI-driven services.

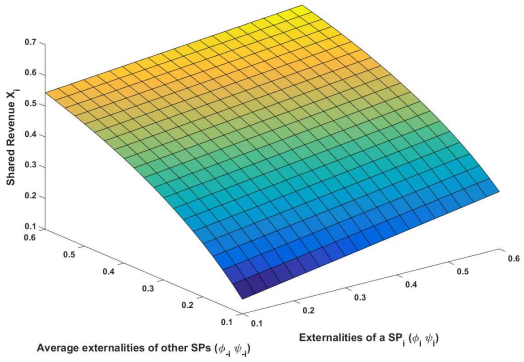


Figure 4.6: Shared revenues X_i over externalities $\phi_i\psi_i$ and $\phi_{-i}\psi_{-i}$

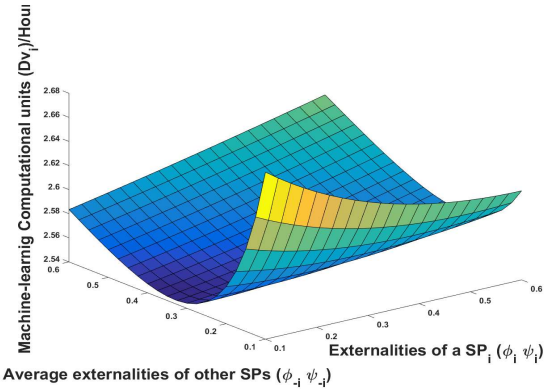


Figure 4.7: Centralized-machine learning computational units Dv_i over externalities $\phi_i\psi_i$ and $\phi_{-i}\psi_{-i}$

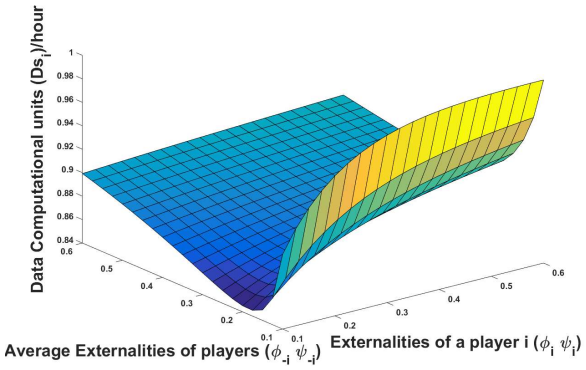


Figure 4.8: Data computational units Ds_i over externalities $\alpha_i\beta_i$ and $\alpha_{-i}\beta_{-i}$

4.4.4 Profitability Analysis

In this section, we study the impact of externalities on the profitability of involved parties, (i.e., data providers, and the cloud computing). As shown in Figure 4.9, the data provider SP_i receives higher net payoff when its externalities $\phi_i\psi_i$ are stronger than other data providers' externalities $\phi_{-i}\psi_{-i}$, (i.e., the service provider SP_i is the most influential among the other data providers). This is due to the low portion of revenues imposed by the cloud computing, as seen previously in Figure 4.6, to subsidize the influencer data provider. However, this is not a case when it comes to the cloud's payoff. As shown in Figure 4.10, the cloud receives higher payoff under strong externalities among data providers. This is interpreted by the high portion of revenues demanded by the cloud from all connected data providers under strong externalities.

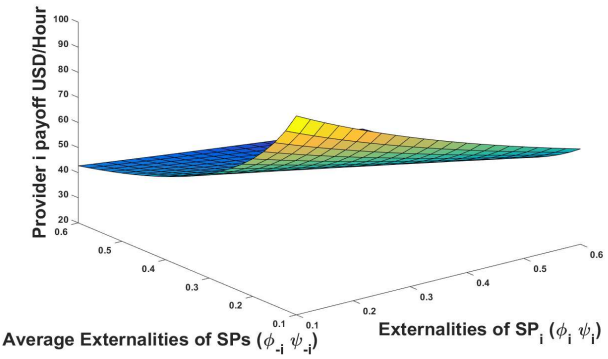


Figure 4.9: Provider $i's$ payoff over externalities $\phi_i\psi_i$ and $\phi_{-i}\psi_{-i}$

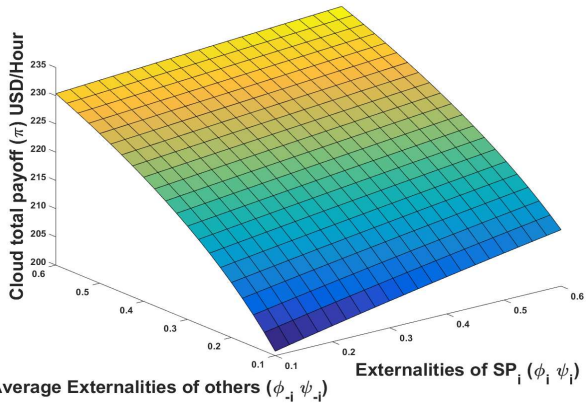


Figure 4.10: Cloud payoff over externalities $\phi_i\psi_i$ and $\phi_{-i}\psi_{-i}$

4.5 Related Work

The business-oriented data trading models can be classified into three categories: merchant, collaborative, the two-sided business model. Under the merchant model, a third-party platform (e.g., information service providers) aims to maximize its revenue through buying data from their owners, reprocessing them, extracting useful information and selling this information to consumers. The interactions among the involved entities, i.e data providers, information provider and consumers, have been implemented using different techniques such as game theory and auctions. This model has been widely applied in domains related to mobile phone sensing networks and IoT services such as [70, 48] and [46]. However, this model is not efficient for the monetization and trading of large-scale of big data commodities. Precisely, this model entails higher processing costs once the consumers exhibit high interests' diversity in terms of extracted information and targeted domains. Moreover, the involved parties in this model are expected to exhibit aggressive competitive behavior in terms of raw data price from one side, and the value of the extracted information from the other side. The resulting equilibrium from such an aggressive competition among the different involved parties leads to less and coarse distribution of the total surplus.

The proposals under the collaborative model tried to mitigate the aggressive competition entailed by the merchant model through establishing collaborative strategies between the data providers and the information service provider. For instance, the authors of [111] and [101], propose game theoretical models among IoT sensors (which act as data providers), IoT service providers (which

act as information service providers) and data consumers. In these games, two entities (i.e., IoT sensors and IoT service providers) cooperate together in one game and then compete as one entity against data consumers. However, those proposal are not able to address how cooperating entities would share their earned revenues due to the absence of the cross-group externailities that help seize the contributions of involved parties.

Recently, the two-sided market model [82] has been proposed as a successful business model to monetize the raw big data over a third-party platform [6, 5, 16]. Under this model, the main role of the third-party platform is introducing a wide social networks of data consumers to the data providers and vice versa. Technically speaking, the data consumers will be attracted to the third-party platform to which a massive number of data providers are connected and vise versa. The strength of attraction between the market sides, i.e., data providers and data consumers has been modeled using the cross group externalities that capture the mutual impact between the providers and consumers. These proposals, i.e [6, 5, 16], focus on answering the following research question: why is it not efficient that the third-party purchases the data from the data providers, reprocesses it, and sells it as information for data consumers when it is applied to big data, (i.e., the merchant model). However, the main limitation of these recent works is that they provide a static analysis of the data prices and consumers' demand for dynamic and active data market environments that are constantly changing such IoT services, and the big data over the cloud computing.

This dynamicity in the data market has been recently addressed in [14, 13] by integrating a mix of cooperative and competitive strategies with the two-sided market theory to monetize the data over the cloud computing and the blockchain technology. Unlike the approaches such as [78, 80, 76] that distribute computing resources of the cloud in its current business version (one-side market), the authors in [14] design a novel model to distribute cloud resources when the cloud computing acts as a third party platform (two-sided market). However, all of the proposals concerning the two-sided data market did not consider the dependencies among the data providers that quantify the monetary value of a certain data type in presence of other different complement data types; complementary data pricing. Moreover, those proposals focus on the traditional data monetizing style that is limited to sell a dataset for data consumers. Such a traditional monetization style is not inline with the modernization of the AI services that enable the execution of the machine learning procedures on

the data without direct access to preserve the privacy of the data users.

Establishing complementary data services has been discussed in many proposals. For instance, the authors of [110] recently propose a game theoretical model to deliver a bundle of complementary IoT services considering the externalities. However, the terms of the externailaties in [110] refers to communication network congestion, see [35, 108, 27]. The solution follows the merchant model scenario. Nevertheless, unlike our solution, these approaches do not target cloud computing. In general, the complementary data services over a third party in presence of the two-sided model has not been addressed yet.

To fill this research gap, we proposed in this paper a novel business game-based model using the two-sided market theory. Our model enables the cloud computing to monetize the complementary data for AI-driven services.

4.6 Conclusion

In this paper, we proposed a novel solution for the challenge of providing complementary data for AI-driven services. The solution mainly sees the cloud computing as an active market platform where data providers and AI-driven services meet each other and exchange mutual benefits in terms of data and monetary reward transactions. Under this version, we proposed a novel two-sided business model for the cloud computing supported by a strategic mechanism to bundle the complementary data. The simulation results showed the efficiency of the proposed model over different ranges of externalities among data providers.

Chapter 5

Trading of Big Data and IoT Services: Blockchain as Two-Sided Market

The blockchain technology has recently proved to be an efficient solution for guaranteeing the security of data transactions in data trading scenarios. The benefits of the blockchain in this domain have been shown to span over several crucial security and privacy aspects such as verifying the identities of data providers, detecting and preventing malicious data consumers, and regulating the trust relationships between the data trading parties. However, the cost and economic aspects of using this solution such as the pricing of mining process have not been addressed yet. In fact, using the blockchain entails high operational costs and puts both the data providers and miners in a continuous dilemma between delivering high-quality security services and adding supplementary costs. In addition, the mining leader requires an efficient mechanism to select the tasks from the mining pool and determine the needed computational resources for each particular task in order to maximize its payoff. Motivated by these two points, we propose in this paper a novel game theoretical model based on the two-sided market approach that exhibits a mix of cooperative and competitive strategies between the (blockchain) miners and data providers. The game helps both the data providers and miners determine the monetary reward and computational resources respectively. Simulations conducted on a real-world dataset show promising potential of the proposed solution in terms of achieving total surpluses for all involved parties, i.e., data providers, data consumers and miners.

5.1 Introduction

Blockchain technology has lately emerged as a revolutionary paradigm for addressing the challenges of finding trustworthy third-parties and guaranteeing the privacy and security of data trading transactaions in critical domains such as Internet of Things (IoT), data analytics, mobile crowdsensing, and machine learning. Interestingly, recent statistics estimate that the data contained in the blockchain ledger is expected to worth up to 20% of the global big data market and to generate up to 100 billion in annual income to the data market that already hit $\$203$ billion dollar of revenue at the end of 2019 [46, 1]. In the context of data trading using blockchain, three players are to be considered: miners, data providers and data consumers. Miners are responsible for supervising and regulating the execution of what is known as *smart contracts*. A Smart contract is a self-executing computer program that states and organizes the agreed terms of a certain data transaction such as the desired quality of service clauses and secure payment mechanism between the data providers and data consumers. Processing smart contacts by miners entails high (mining) operational costs and processing time, which might negatively affect the execution time of real-time and delay-critical applications such as IoT and data analytics. In the literature, there is lack of attention on the business model that would enable data trading over blockchain where the main stream research in the general context of data focuses on developing mechanisms of data resource management such as [79, 80, 78]. Several challenging issues are yet to be addressed, in particular, assigning optimal amount of computational units to the mining tasks, sustaining optimal payoffs to involved players and serving data requests on time. In this work, our objective is to provide a novel contribution to the data trading over blockchain through proposing a game-theoretic-based business model that helps regulate the secure data trading of IoT and big data analytics services. In particular, we aim to address the following two substantial research challenges: 1) how should the blockchain node distribute the computational resources of the mining process among the data providers in such a way to maximize its payoff; and 2) how should the data providers decide on the optimal monetary reward that needs to be given to the miners versus their service in such a way to guarantee optimal execution time of their transactions while avoiding over-payments.

5.1.1 Motivating Example

We provide in Figure 5.1 a motivating example to better clarify the research gap in the literature and highlight the need of our solution. As explained in the figure, data consumers request to run real-time data analytics on an edge IoT server. Following the blockchain technology, the request is deployed as a smart contract which includes clauses that regulate the relationships between the data consumers and the edge IoT server in terms of data quality, data size and processing speed. The execution of the smart contract is supervised and executed by the blockchain node, which manages the mining process and the mining computational units. Smart contracts vary in their terms, and hence they differ in their executions in terms of execution time and required resources. For instance, in Figure 5.1, the hospital server is exposed to more privacy threatens as it stores patients data, which requires more computational units from the blockchain node to authenticate only trusted consumers. This creates the need for a distributing mechanism that determines the optimal amount of resources for each smart contract. However, the absence of such a mechanism might assign more resources to less profitable contracts.

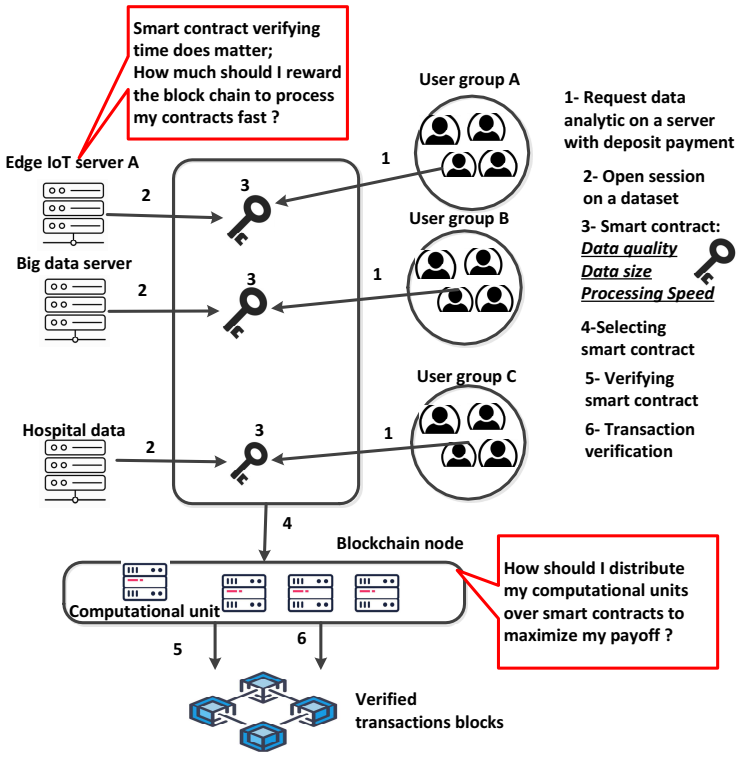


Figure 5.1: Motivating Scenario: Run real time data analytics procedures on Edge IoT server using the blockchain technology.

5.1.2 Related Work and Problem Statement

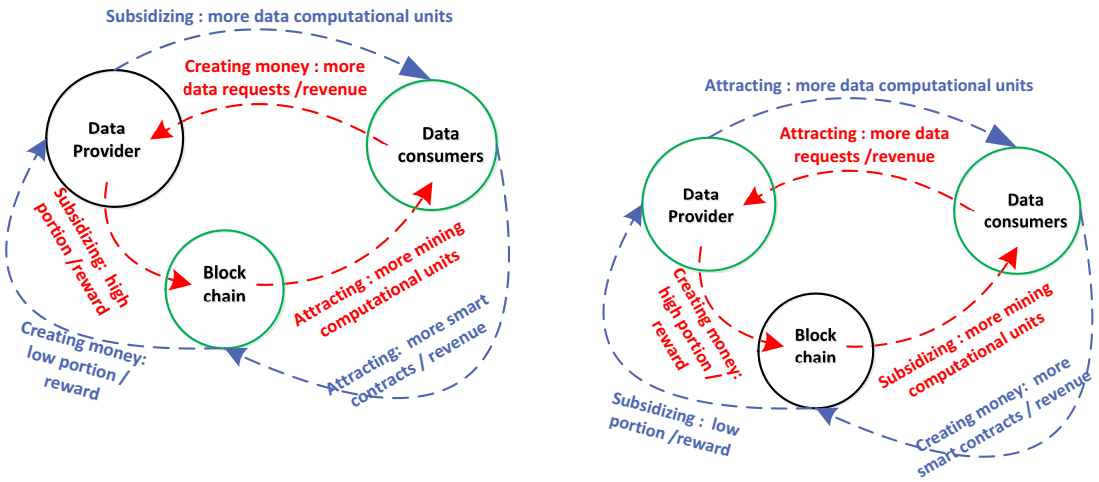
The state-of-the-art proposals focus on deploying verification approaches into the blockchain technology in order to tackle the privacy and security issues such as preserving the anonymity of the data providers, and preventing impersonation attacks and colluding miners. For instance, the approaches proposed in [98, 105] leverage the blockchain technology to address the problem of user location impersonation and re-identification attacks respectively in a crowd-sensing context. The approaches proposed in [42, 50] aim to increase the engagement of the crowd system participants through capitalizing on the anonymous and reliable interaction features provided by the blockchain technology.

The proposals [100, 47, 61, 99] propose game theoretical foundations in the context of mobile blockchain supported by edge computing services. The interactions between miners and edge computing nodes are modeled using Stackelberg games and auctions to derive an optimal price for the proof-of-work for offloading allocation tasks. The main limitation of such games is that they result in putting the miners into an aggressive competition situation between each other from one side, and with the edge computing services from the other side. This leads to less efficient outcomes in terms of total surpluses for all these parties. In [102], the authors propose to deploy blockchain for big data sharing in a collaborative edge environment. Similar works have also been proposed in [58, 107]. The aforementioned proposals, and the state-of-the-art in general suffer from several problems. In fact, they 1) do not explain how the mining resources should be distributed over the existing smart contracts and miners; 2) do not provide any mechanism to derive the optimal payment that should be given by data providers to miners); and/or 3) propose pricing schemes for the mining process based on pure competitive games, which entails an aggressive competition among the involved parties and results in lower payoffs for them.

5.1.3 Contributions

To address the aforementioned issues, we extend the work in [17, 16] by proposing a novel double two-sided game that models the interactions among the involved parties (i.e., blockchain node, data providers and data consumers) using the two-sided market theory [82]. In the proposed game, as shown in Figure 5.2, both the data providers and blockchain node act as a two-sided platform that gets on board two market sides. Specifically, the blockchain node intermediates the

interactions between the data providers and data consumers, while the data providers intermediate the interactions between the blockchain node and data consumers. As shown in the figure, the data providers either 1) subsidize the blockchain node by a higher portion of revenue to motivate it to supply more mining computational units, which results in attracting more data consumers and increasing the revenue; or 2) subsidize the data consumers by more data computational units, which increases the consumers' demand and hence contributes in attracting the blockchain node. Similar strategies are set up to the blockchain node as shown in Figure 5.2b. The proposed game combines both strategies as two separate games. The solution of the games helps derive the equilibrium in terms of shared revenue among the blockchain node and data providers and amount of mining resources that each smart contract should be assigned with.



(a) Data provider as two-sided market

(b) Blockchain as two-sided market

Figure 5.2: Proposed model: A double two-sided market game

5.2 Proposed Model for Secure Trading of Data

5.2.1 Model Description: A Double Two-Sided Game Formulation

The proposed secure data trading model, depicted in Figure 5.3, consists of three entities: Data Service Consumers (SC), Big Data Service Providers (SP) and Blockchain node (BC) that consists of a network of miners. In our solution, a certain big data service provider SP_i receives a monetary

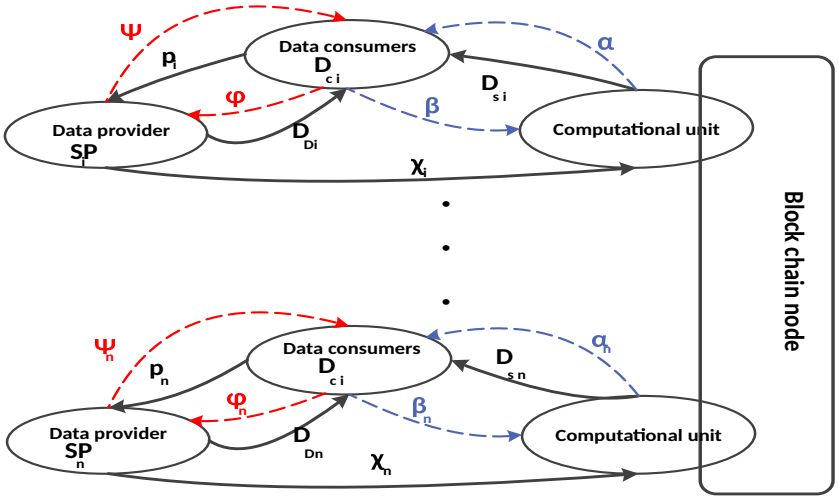


Figure 5.3: Double two-sided game

value of P_i per a data service consumer's access to its services. The service provider SP_i provides both the data and computing resources that are required to execute the data analytics duties of the data consumers. The interactions between data providers and data consumers include negotiating the data type, quality of provided services, payments, and all the associated terms of delivered data services. The blockchain node BC is in charge of executing the transactions of smart contracts in order to append a correct block into the blockchain. Executing smart contracts will also ensure the sustainability of consumers' access security, verification of the identities of the data providers and consumers, protection of the privacy of data providers and enforcement of quality control of data services. In our model, the blockchain node seeks to distribute and allocate its computing resources for the mining process among service providers in such a way to maximize its own payoff.

The Consumers' demand on data service i provided by a service provider SP_i is denoted by D_{c_i} and the computing resources allocated by service i to run the data analytics duties of its consumers is donated by D_{D_i}. D_{D_i} is measured in terms of the throughput per second of executing the data requests. The relationship between consumers' demand D_{c_i} and supplying service i is modeled using the two-sided market theory [82] as cross-group externalities ϕ and ψ. Here, ψ represents the increase in the number of data consumers obtained when some new computing and storage resources are added to D_{D_i}. ϕ represents the amount of profit that the data service provider earns when one more new consumer is added to D_{c_i}. Similarly, the computing resources allocated by the blockchain node to regulate the smart contracts of service i is denoted by D_{s_i}. The relationship

between consumers' demand and the supply of the blockchain node is likewise modeled using the two-sided market theory as cross-group externalities α and β. Here, α represents the increasing of data consumers obtained when some new computing and storage resources are added to D_{s_i} and β represents the amount of benefits that the blockchain node earns when one more new consumers are added to D_{c_i}. The parameters α, β, ϕ, and ψ are dependant on the service i. However, the variable i is omitted from the notations of these parameters to simplify the equations when the service i is understood from the context. Thus, instead of using α_i for instance, α will be used. The same simplification is applied for the other parameters that appear as exponents in our equations.

The interaction between SP and BC is modeled as a two-stage game, where BC acts as the game leader and SP are the followers. In the first stage of the game, each service provider SP_i providing service i observes the amount of money returns χ_i requested by BC in order to adjust the supply volume of computing resources and the price to be charged to service consumers SC_i consuming the service i. In quest of the price specified by SP_i, BC determines the optimal amount of computing resources D_{s_i} that should be supplied to handle the smart contracts between SP_i and SC_i. The model forms a closed loop of dependencies that involves subsidizing techniques from the two-sided market theory. Thus, SP_i may chose to subsidize BC by an extra amount of payment that exceeds the contribution of BC. The objective is to keep an optimal level of D_{s_i} that maximizes the return revenues $P_i * D_{c_i}$. Alternatively, BC may subsidize SP_i with a low portion of the resulting revenue to keep an optimal level of P_i. The different parameters and symbols used in our proposed solution are summarized in Table 5.1.

5.2.2 Players Demands and Utility Functions

The consumer's demand and supply are modeled using the Cobb-Douglas function, which have the ability to represent the elasticity of the computing and storage resources supply (D_{s_i}, D_{D_i}) and its variations depending on the user's demand. These demand functions are defined as per Equations (70), (71), and (72). By substituting Equations (71) and (72) into Equation (70), we can express the consumer's demand as a function of χ_i and P_i as described in Equation (73).

$$D_{c_i} = k_1 P_i^{-\gamma} D_{s_i}^{\alpha} D_{D_i}^{\psi} \tag{70}$$

Model Parameters	Descriptions.
SP_i	Service provider providing service i.
BC	A blockchain node.
SC_i	Consumers of service i.
D_{c_i}	SC_i's demand.
D_{D_i}	IT-infrastructure supply to handle requests of SC_i.
D_{s_i}	IT-infrastructure supply to handle smart contracts between SP_i and SC_i.
P_i	Service i's price.
χ_i	Portion of revenue required by BC from SP_i.
α	The Network effects (externality) on D_{c_i} by D_{s_i}.
β	The Network effects (externality) on D_{s_i} by D_{c_i}
ψ	The Network effects (externality) on D_{c_i} by D_{D_i}.
ϕ	The Network effects (externality) on D_{s_i} by D_{c_i}.
γ	D_{c_i}'s elasticity with respect to P_i.
k_1, k_2, and k_3	Constant multipliers.
f_c	Associated costs per smart contract.
f_s	Associated costs per service request by a consumer.
π_i	SP_i's payoff.
π	Blockchain node's payoff.
a_1	$= -\gamma + \alpha\beta + \phi\psi$
a_2	$= \alpha\beta$
a_3	$= 1/(1 - \alpha\beta - \psi\phi)$

Table 5.1: Model parameters

$$D_{s_i} = k_2(\chi_i P_i D_{c_i})^{\beta} \tag{71}$$

$$D_{D_i} = k_3(P_i D_{c_i})^{\phi} \tag{72}$$

$$D_{c_i} = (k_1 k_2^{\alpha} k_3^{\psi} P_i^{a_1} \chi_i^{a_2})^{a_3} \tag{73}$$

Each big data service provider SP_i is subject to a fixed cost f_s per each consumer access. SP_i aims to maximize its payoff as described in Equation (74).

$$\pi_i = ((P_i)(1 - \chi_i) - f_s)D_{c_i} \tag{74}$$

The blockchain node BC is subject to a fixed cost f_c per each smart contract between SP_i and a data consumer. As a rational agent, the blockchain node seeks to maximize its payoff as given in

Equation (75).

$$\pi = (P_i \chi_i - f_c) D_{c_i} \tag{75}$$

5.2.3 Game Equilibrium

The equilibrium of the above-described game is solved using a backward induction methodology. Specifically, the followers' (data service providers) sub-game is solved first to obtain their optimal response P_i^* to the service consumers. The leader's (blockchain node) sub-game is then computed to obtain the optimal χ_i^*. The game equilibrium is stated in Theorem 4.

Theorem 4. *Under the assumption validated in [16] stating that the cross-group externalities are not too week and not too strong,* $(0.1 < \alpha\beta < 0.8)$ *and* $(0.1 < \phi\psi < 0.8)$*, The equilibrium of our double two-sided game is given by the best responses of the different players as follows:*

(1) The best response of the data service provider SP_i *is given by:*

$$P_i^* = \frac{a_1 a_3 f_s}{(a_1 a_3 - 1)(\chi_i^* - 1)} \tag{76}$$

if: $\quad 1 < (1/a_1 a_3)$

(2) The best response of the Blockchain node with respect to a service i *is given by:*

$$\chi_i^* = \frac{a_2 a_3 f_c}{(a_2 a_3 + 1) P_i^*} \tag{77}$$

Proof. From Equation (74) of the data service provider's payoff, using $\log$ for both sides of the equation, we obtain:

$$\log \pi_i = \log(P_i(1 - \chi_i) - f_s) + \log D_{c_i} \tag{78}$$

Then, the optimal price P_i^* is defined by $\partial \pi_i / \partial P_i = 0$ as follows:

$$\frac{1}{\pi_i} \times \frac{\partial \pi_i}{\partial P_i} = \frac{1 - \chi_i}{P_i(1 - \chi_i) - f_s} + \frac{1}{D_{c_i}} \times \frac{\partial D_{c_i}}{\partial P_i} = 0 \tag{79}$$

By deriving Equation (73) with respect to P_i, then:

$$\frac{\partial D_{c_i}}{\partial P_i} = a_1 a_3 D_{c_i} P_i^{-1} \tag{80}$$

By substituting Equation (80) into Equation (79), we get:

$$P_i = \frac{a_1 a_3 f_s}{(a_1 a_3 - 1)(\chi_i - 1)} \tag{81}$$

Since $P_i > 0$, $fs > 0$, $((\chi_i - 1) < 1)$ then $(a_1 a_3/(a_1 a_3 - 1) < 0)$, so the condition. By considering the acceptable range for γ analysed in [31], $0.2 < \gamma < 0.3$ then $\partial \pi_i / \partial P_i > 0$ when $P_i < (a_1 a_3 f_s)/((a_1 a_3 - 1)(\chi_i - 1))$ and $\partial \pi_i / \partial P_i < 0$ when $P_i > (a_1 a_3 f_s)/((a_1 a_3 - 1)(\chi_i - 1))$. Consequently, P_i is the best response.

For the second result of the theorem, we consider and take the log for both sides of the equation of the blockchain node's payoff (Equation (75)) and obtain:

$$\log \pi = \log(P_i \chi_i - f_c) + \log D_{c_i} \tag{82}$$

Then, the optimal χ_i^* is defined by $\partial \pi / \partial \chi_i = 0$ as follows:

$$\frac{1}{\pi} \times \frac{\partial \pi}{\partial \chi_i} = \frac{P_i}{P_i \chi_i - f_c} + \frac{1}{D_{c_i}} \times \frac{\partial D_{c_i}}{\partial \chi_i} = 0 \tag{83}$$

By deriving Equation (73) with respect to χ_i, we get:

$$\frac{\partial D_{c_i}}{\partial \chi_i} = a_2 a_3 D_{c_i} \chi_i^{-1} \tag{84}$$

By substituting Equation (84) into Equation (83), then:

$$\chi_i = \frac{a_2 a_3 f_c}{(a_2 a_3 + 1) P_i} \tag{85}$$

$\partial \pi / \partial \chi_i > 0$ when $\chi_i < (a_2 a_3 f_c)((a_2 a_3 + 1) P_i)$ and $\partial \pi / \partial \chi_i < 0$ when $\chi_i > (a_2 a_3 f_c)((a_2 a_3 + 1) P_i)$. Consequently, χ_i is the best response, so the theorem.

□

5.3 Simulation and Empirical Analysis

5.3.1 Simulation Setup

Our simulation analysis is grounded on statistical observations from big data and IoT services from the AWS marketplace [9], BMR [2]—the annual statistical report that publishes the revenues, payoffs and market growth of the the AWS marketplace—and a real-world dataset from Google [37]. The price, P_i, of the data service is chosen from the interval $[0.2, 3.2]$ USD/hour, following

the price distribution of 150 data and IoT services from the AWS marketplace. According to [2], Amazon Web services (AWS) received 30 billion USD in revenue with a net income of approx. 12 billion. The gap between the gross and net revenues is caused by the marginal operating costs which made up approx. 60% of revenue. The operating costs represents in our model the costs associated with the smart contracts f_c and service requests initiated by data consumers f_s. The Google dataset [37] records statistics on the execution of big data requests executed on Google-powered virtual machines, which are similar to the instances of Amazon cloud infrastructure (EC2). According to these statistics, each virtual machine takes on average 1.42 to 10 seconds to complete a data processing request (with a mean of 5.71s and standard deviation of 4.29s). The instances and their average computational power are respectively represented in our model by D_{s_i} and the externality factor α. Adding a compute instance has a direct impact on the increase of the consumers' demand between 0.1 to 0.7 data request per second. By following the mathematically proved result in [16] that the cross-group externalities should not be neither too weak nor too strong, the cross-group externalities should be bounded by $0.1 < \alpha\beta < 0.8$. Hence, the externality factor β would range from $0.1/\alpha$ to $0.8/\alpha$. We follow those estimations and set up the cross-group externalities ϕ and ψ in the same range of α and β. The price elasticity γ is set to 0.15, which is similar to the sensitivity of mobile/telecommunication services price estimated in the literature [31]. The simulation takes the aforementioned parameters as inputs, and then calculates the optimal shared revenue χ_i from each service i according to Equation (77) in Theorem 4. Moreover, the simulation inputs meet the theoretical condition $(1 < 1/a_1a_3)$ in Theorem 4. Thus, by substituting the real ranges of the simulation parameters, the mathematical term representing the strength of total externalities (a_3) is greater than zero (i.e $\alpha\beta + \phi\psi < 1$). Hence, we demonstrate our three dimensional results in three sets of criteria: 1) week externalities ($0.1 < \alpha\beta < 0.4$, $0.1 < \phi\psi < 0.4$); 2) strong externalities of $\alpha\beta$ - weak externalities of $\phi\psi$ ($0.4 < \alpha\beta < 0.7$, $0.1 < \phi\psi < 0.2$); and 3) strong externalities of $\phi\psi$ - weak externalities of $\alpha\beta$ ($0.1 < \alpha\beta < 0.2$, $0.4 < \phi\psi < 0.7$).

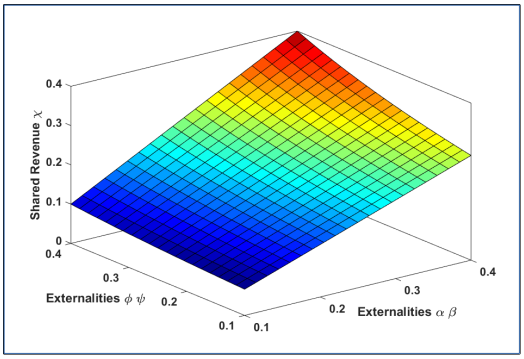


Figure 5.4: Shared revenue over week externalities

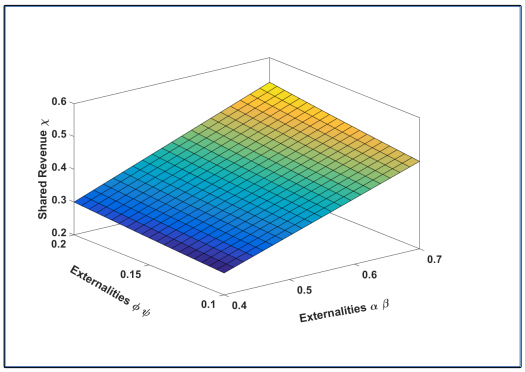


Figure 5.5: Shared revenue over strong externalities $\alpha\beta$

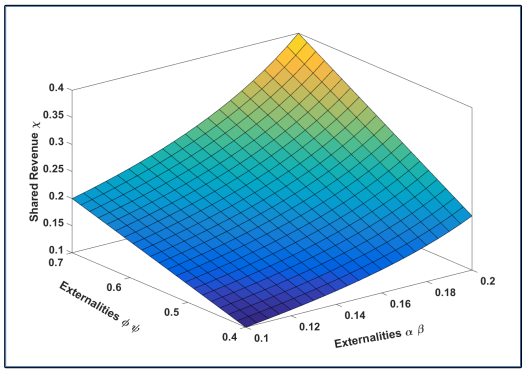


Figure 5.6: Shared revenue over strong externalities $\phi\psi$

5.3.2 Shared Revenues and Computational Costs over Externalities

In this section, we study the impact of the cross group externalitie metrics $(\alpha\beta)$ and $(\phi\psi)$ on the shared revenue χ_i among data providers and blockchain node. In Figure 5.4, we study the percentage of shared revenue received by the blockchain node for a weak level of externalities between, on one side the data providers and blockchain node, and on the other side the data consumers. In Figures 5.5 and 5.6, we study the shared revenue for a stronger level of externalities $\alpha\beta$ and $\phi\psi$ respectively. As shown in these figures, the blockchain node receives a higher percentage of revenue as the externality factors $\alpha\beta$ and $\phi\psi$ become stronger. Another important observation is that the average of shared revenues increases at a higher pace over the blockchain node externalities with data consumers $(\alpha\beta)$ than that over data provider and data consumers $(\phi\psi)$. This behavior is clearly observed in Figure 5.5 which shows that the shared revenue reaches 60% over strong externalities of $\alpha\beta$ versus a maximum of 40% over strong externalities of $\phi\psi$ as shown in Figure 5.6. This behavior is interpreted as follows. The demand of data consumers is positively impacted when

its externalities with the blockchain node ($\alpha\beta$) become stronger. Consequently, the data providers entice the blockchain node by a higher portion χ_i of revenues to supply more computational units with the aim of increasing the consumers' demand and hence the total revenue. Nonetheless, the blockchain node faces higher operating costs by increasing its supply of mining computational units. Consequently, it would ask for a higher portion of revenue. Moreover, the consumers' demand is positively impacted as its externalities with data provider become stronger. Thus, the data providers would face higher operating costs when they add more computational units in an attempt to increase the consumers' demand. This forces the blockchain node to subsidize data providers with a lower portion χ_i of revenue to sustain a higher level D_{c_i} of consumers' demand. In general, increasing the consumers' demand adds more computational cost on the blockchain node, which leads to increasing the portion of blockchain node as the externalities among the data provider and data consumers become stronger. This explains the slower increase pace of shared revenues over the externalities $\phi\psi$ compared to the externalities $\alpha\beta$.

5.3.3 Data Consumers' Demand and Computational Unit Supply

In this section, we study the impact of cross-group externalities among all the involved parties (i.e., data providers, blockchain node, and data consumers) on the data consumers' demand. As shown on Figures 5.7, 5.8 and 5.9, the consumers' demand is higher under a weak level of externalities than the strong level. Those observed results are interpreted as follows. A higher externality level among the market players incurs a higher cost for the two-sided market platform to get the market players on board. Specifically, under a strong level of externalities among the blockchain node and data consumers $\alpha\beta$, data providers either (1) subsidize the blockchain node with a higher portion of revenue to attract more data consumers (as discussed in Section 5.3.2); or (2) subsidize the data consumers by supplying higher amounts of data computational units, which in turns, leads to incentivizing the blockchain node. However, data providers cannot ultimately subsidize data consumers due to their mutual cross-group externalities ($\phi\psi$). To study this phenomenon, we show in Figures 5.10 and 5.11 the amount of data computational units supplied by data providers as well as the number of data consumers attracted over the externalities $\phi\psi$ respectively. As shown in Figure 5.10, the amount of supplied computational units increases under weak externalities ($\phi\psi \in [0.1 - 0.4]$) and

gradually decreases as the cross-group externalities become stronger (i.e., $\phi\psi \in [0.4 - 0.8]$). However, as shown in Figure 5.11, the number of attracted data consumers exponentially decreases over the whole range of externalities. This implies that the subsidizing technique becomes costly as the externalities become stronger. For instance, data providers attract 2×10^5 data consumers by providing 20 data computational units at an externality level of 0.2, while they attract a number of data consumers that is 0.1×10^5 less by providing the same amount of data computational units but with a higher externality level of 0.5. In both cases (i.e., subsidizing data consumers and data providers), the data providers would undergo higher costs. Similarly, under a strong level of externalities between data providers and data consumers, the blockchain node subsidizes either the data providers (by asking lower portion of revenues) or the data consumers (by supplying a higher amount of computational units), which entails higher costs for both cases. Similarly, the blockchain node cannot ultimately subsidize the data consumers due to their mutual cross-group externalities represented by $\alpha\beta$. Similar observations are depicted in Figure 5.12 in terms of mining computational units over $\alpha\beta$.

5.3.4 Data Providers and Blockchain Payoffs

In this section, we investigate the impact of externalities on the payoff of the data providers and blockchain node. Figure 5.13 shows the payoff of data providers under weak externalities, while Figures 5.14 and 5.15 depict providers' payoff under strong externalities $\alpha\beta$ and $\phi\psi$ respectively. As illustrated in these figures, the data providers' payoff gradually decreases as the externalities increase. The reason behind this increasing is that the overall demand of consumers decreases while computational costs and asked shared revenue increase over externalities as discussed in Sections 5.3.2 and 5.3.3. Similarly, the payoff of the blockchain node decreases under externalities as shown Figures 5.16, 5.17 and 5.18.

5.4 Conclusion

In this work, we proposed a new game-based business model for data trading over blockchain. The problem is formulated as a double two-sided game that solved the problem of maximizing the players' payoff by optimally distributing the mining computational powers over smart contracts.

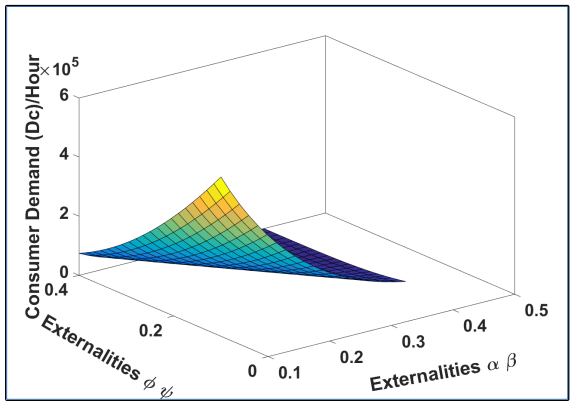


Figure 5.7: Consumers' demand over week externalities

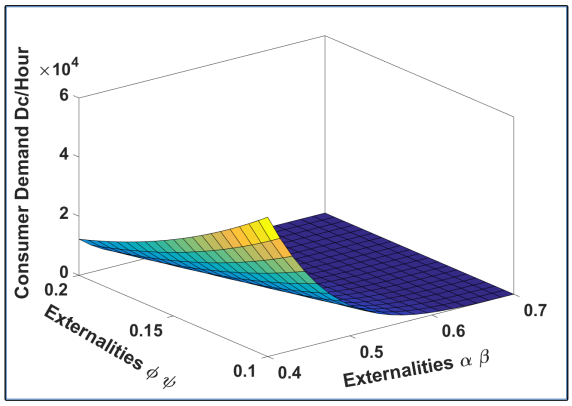


Figure 5.8: Consumers' demand over strong externalities $\alpha\beta$

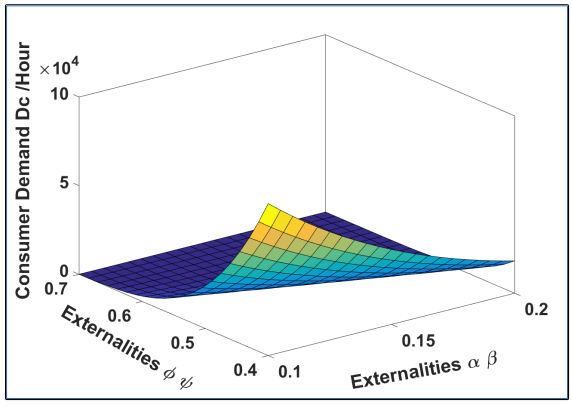


Figure 5.9: Consumers' demand over strong externalities $\phi\psi$

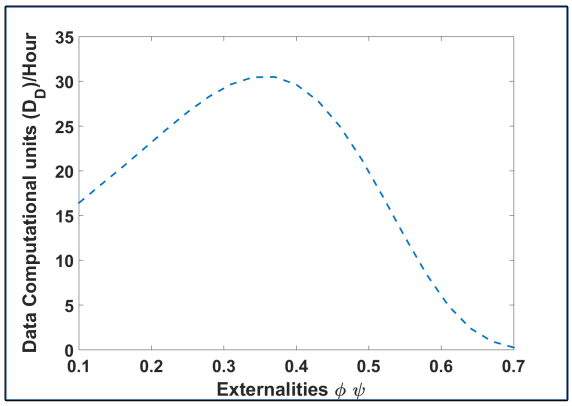


Figure 5.10: Data computational units over $\phi\psi$

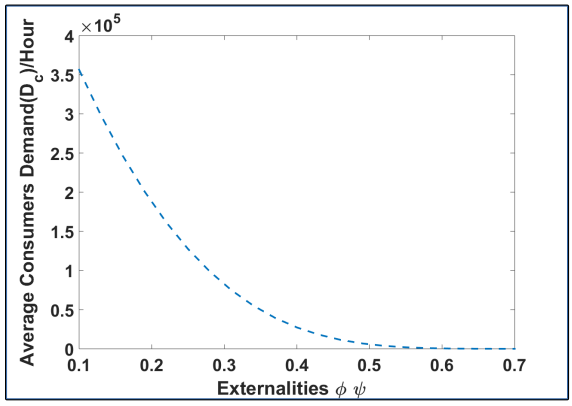


Figure 5.11: Number of attracted consumers over $\phi\psi$

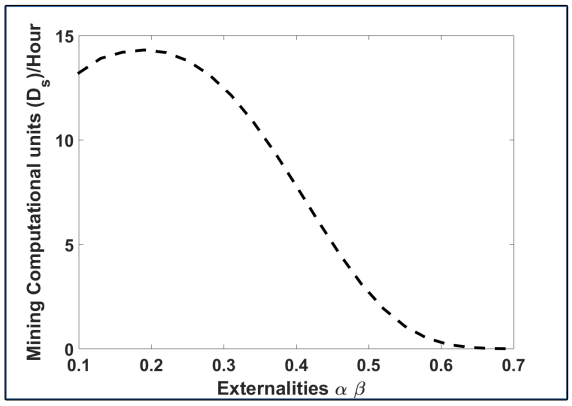


Figure 5.12: Mining computational units over $\alpha\beta$

Technically, the game considered the smart contract characteristics as well as the impact of the mining computational units on the data service and consumers' demand. The theoretical and simulation

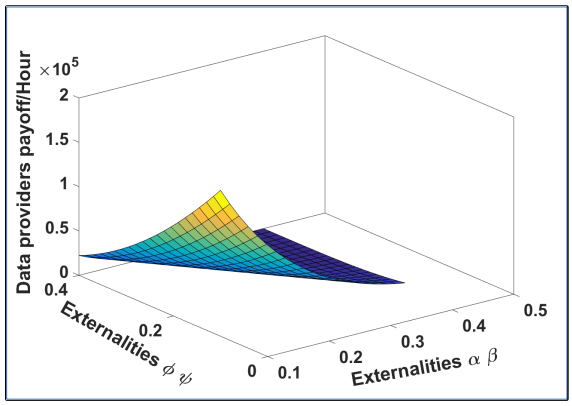


Figure 5.13: Data providers payoff over weak externalities

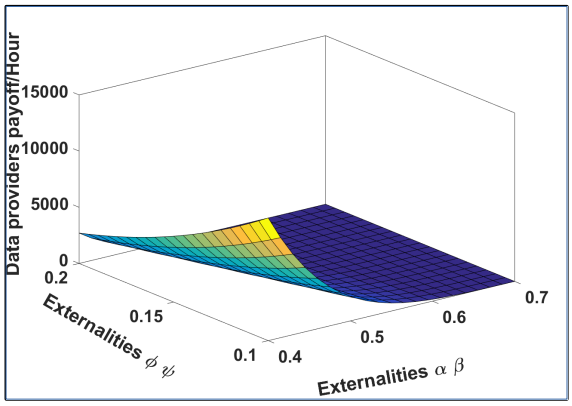


Figure 5.14: Data providers payoff over strong externalities $\alpha\beta$

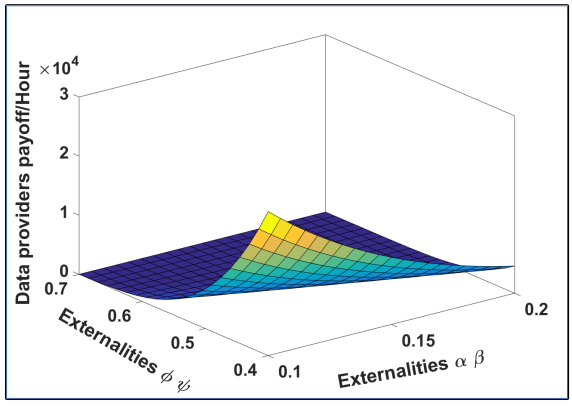


Figure 5.15: Data providers payoff over strong externalities $\phi\psi$

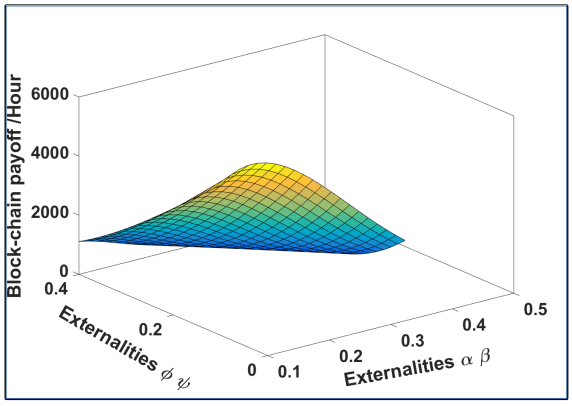


Figure 5.16: Blockchain payoff over weak externalities

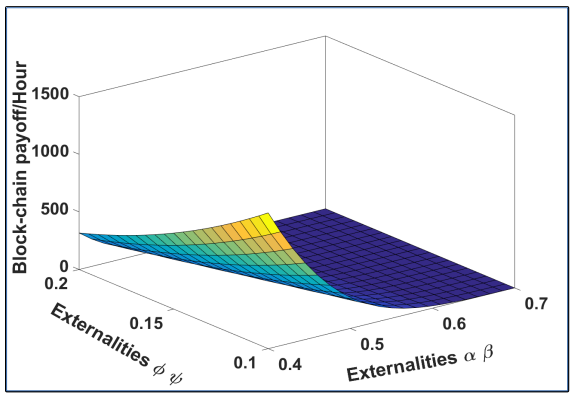


Figure 5.17: Blockchain payoff over strong externalities $\alpha\beta$

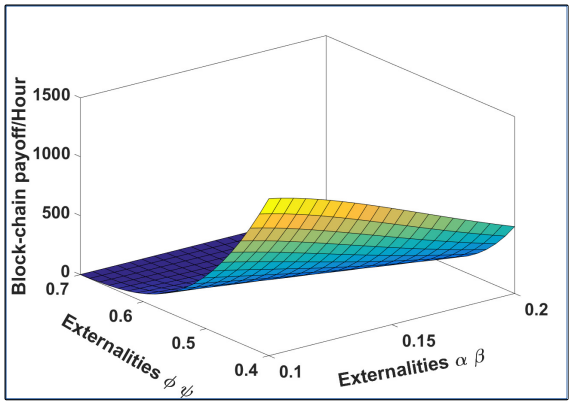


Figure 5.18: Blockchain payoff over strong externalities $\phi\psi$

results proved the efficiency of the proposed game.

Chapter 6

Conclusions and Future Work

6.1 Conclusions and Discussion

This thesis tackled the problem of designing a data marketplace over the cloud computing technology. The problem was discussed from two points of view: 1) What is most efficient business model to design the data market platform? And 2) what is the most efficient technology that can act as data market platform?. As answer to the first question, this thesis proved theoretically and empirically that the two sided market theory is an efficient solution to model the data marketplace. In fact, the merchant and peer to peer models, the most popular models in the literature that concern trading data, raise serious issues in terms of involving the actual data owners and the aggressive competition among involved parties. This competition leads to low social welfare and coarse distribution for revenues. Besides these aforementioned drawbacks, those models did not provide an efficient solution for the fundamental problems in this context: 1) Where data consumers should go to find big data providers that meet the required quality and quantity in a timely manner? 2) Where data providers should go to meet large number of data consumers to achieve efficient revenues? And 3) how much is the truly price of the data?

These fundamental problems appeared earlier in the area of mobile phone and sensing network applications and caused a turn down in those applications. Mainly, mobile phone applications (data consumers) entailed high costs and management obstacles associated with finding and collecting data. From another perspective, mobile phone users (data providers) exhibited limited willingness to be involved in these applications due to less revenue resulting form reaching a few number of data

consumers. These problems would have been solved if there had been found a data market where data providers and data consumers would meet. These problems have re-appeared recently in the era of AI-driven services and machine learning applications. Consequently, scientific research communities expect a turn down in the revolution of AI-driven services. In this thesis, we propose and analyze a novel platform for data monetization using two-sided market theory. The proposed platform is a coordinated marketplace that facilitates the search for data providers and data consumers and allow them to meet to exchange financial benefits. The proposed platform provides a solution for using personal data by involving individuals in the data monetization process. Furthermore, the platform helps data consumers increase the engagement of data providers, and get easy and quick access to high quality personal data. Consequently, 1) searching costs have been cut dramatically; and 2) user's privacy has been boosted by giving individuals the control of with whom they share specific personal data and get proper compensation.

The second question discussed in the thesis is: what is the most efficient technology that can act as data market platform? This thesis argued that cloud computing is the answer. Mainly, the cloud hosts an explosive amount of data coming from a variety of enterprises and manufacturers that are deployed on its computing platforms. In addition, the cloud computing is equipped by a powerful IT infrastructure including servers and virtual machines to run and deploy computational tasks associated with storing and processing the data. Our two-sided cloud computing platform for monetizing data raises challenges associated with distributing elastically the cloud computing resources and revenue maximization. This platform allows three entities to interact: cloud computing, data providers, and data consumers. Moreover, AI-driven services and data collecting-based applications (data consumers) require big chunks of complementary dataset coming from multiple data providers. The data types in complementary datasets exhibit a range of correlations and dependencies in the sense that the availability of a certain data type impacts the monetary value of other data types. In this thesis, we proposed a game theoretical model based on the two-sided market theory to monetize complementary big data for AI-driven services over the cloud. The objective is to come up with a new vision in which the cloud can play a primordial role in introducing big data service providers and data consumers to each other, which results in higher benefits for all the involved players. The elasticity of the cloud computing is integrated with the two-sided market theory to

maximize the profit of data providers, data consumers and cloud computing, and distribute dynamically the cloud resources among the computational tasks. The proposed game model comes up with a solution for uncertain externalities in the two sided market model. Moreover, the proposed game model provides a clear and efficient mechanism to split the revenues among the involved parties, which is not addressed by the corresponding collaborative models in the literature. This model is supported by a strategic mechanism to bundle the complementary data. Empirical results showed that our solution outperforms the state-of-the-art cloud business models, i.e., the egalitarian and pay-as-you-go models in terms of total surpluses earned by the different parties. The simulation results also showed the efficiency of the proposed model over different ranges of externalities among data providers.

The thesis involved the blockchain technology in the process of monetizing data. The blockchain technology has recently proved to be an efficient solution for guaranteeing the security of data transactions in data trading scenarios. This thesis argued that there is lack of attention on the business model that would enable data trading over blockchain. In particular, the mining leader requires an efficient mechanism to select the tasks from the mining pool and determine the needed computational resources for each particular task in order to maximize its payoff. In this thesis, we proposed a new game-based business model for data trading over blockchain. The problem is formulated as a double two-sided game that solved the problem of maximizing the players' payoff by optimally distributing the mining computational powers over smart contracts. Technically, the game considered the smart contract characteristics as well as the impact of the mining computational units on the data service and consumers' demand. The theoretical and simulation results proved the efficiency of the proposed game.

6.2 Future Work

As future work, the model can be extended to involve multiple cloud providers. This step is challenging and is worth being handled in a separate paper. Specifically, the following research questions need to be answered when dealing with such a scenario.

(1) Can the data providers deploy their data over only one cloud provider or do they need multiple

cloud providers in parallel? This scenario is known as multi-homing and single homing in the literature of the two-sided market theory.

(2) Can data consumers access only one cloud provider or multiple cloud providers at a time? This question largely affects the strategies of the cloud providers.

(3) Are the cloud providers in a pure competitive mode or can they collaborate together to deliver, for example, complementary data or execute a certain data analytic service? If they are in pure competitive mode, how will the imperfect information about each other be handled? We argue that this can be solved by using an intelligent prediction mechanism based on machine learning.

(4) What are the preferences of the data providers? As known in practice, the quality and price offered by the cloud providers are not the only factors in this scenario. The trust and reputation of cloud computing also play a critical role.

(5) What are the preferences of data consumers? Are data consumers interested in specific data providers?

As another potential future work, the model can be integrated in an immune system to fight the security attacks that try to manipulate either the data itself or the machine learning techniques to produce misleading results. This brings us to the field of adversarial machine learning which lies at the intersection of computer security and machine learning. This field focuses on enabling the safe adoption of machine learning techniques by ensuring that the inputs to the machine learning algorithms have not been intentionally manipulated by some attackers to cause the training model to produce erroneous results.

Appendix A

My Appendix

A.1 Proof of Proposition 1

A.1.1 Proof of Proposition 1.1

Consider the demand Equations 73 and 8, first derivatives and demand's elasticities with respect to transaction fees are given as follows:

$$\frac{1}{N_{c|q}} \times \frac{\partial N_{c|q}}{\partial p_c} = \frac{\beta_c}{(\alpha_c\alpha_s - 1)p_c}, \qquad \frac{1}{N_{s|q}} \times \frac{\partial N_{s|q}}{\partial p_c} = \frac{\alpha_s\beta_c}{(\alpha_c\alpha_s - 1)p_c} \quad \Rightarrow \quad \frac{\partial N_{s|q}/\partial p_c}{N_{s|q}} = \alpha_s \times \left(\frac{\partial N_{c|q}/\partial p_c}{N_{c|q}}\right) \tag{86}$$

$$\frac{\partial N_{s|q}/\partial p_c}{N_{s|q}} = \alpha_s \times \left(\frac{\partial N_{c|q}/\partial p_c}{N_{c|q}}\right) \quad \Rightarrow \quad \eta^p_{p_c} = \alpha_s \eta^c_{p_c} \tag{87}$$

$$\frac{1}{N_{s|q}} \times \frac{\partial N_{s|q}}{\partial p_s} = \frac{\beta_s}{(\alpha_c\alpha_s - 1)p_s}, \qquad \frac{1}{N_{c|q}} \times \frac{\partial N_{c|q}}{\partial p_s} = \frac{\alpha_c\beta_s}{(\alpha_c\alpha_s - 1)p_s} \quad \Rightarrow \quad \frac{\partial N_{c|q}/\partial p_s}{N_{c|q}} = \alpha_c \times \left(\frac{\partial N_{s|q}/\partial p_s}{N_{s|q}}\right) \tag{88}$$

$$\frac{\partial N_{c|q}/\partial p_s}{N_{c|q}} = \alpha_c \times \left(\frac{\partial N_{s|q}/\partial p_s}{N_{s|q}}\right) \Rightarrow \eta^c_{p_s} = \alpha_c \eta^p_{p_s} \tag{89}$$

where p_c, p_s, and $\alpha_c\alpha_s$ are different from zero. We rewrite platform payoff given by Equation 11 by taking the logarithm for its sides as follows:

$$\log \pi = \log \gamma + \log(p_c + p_s - f) + \log N_{c|q} + \log N_{s|q} \tag{90}$$

By deriving the platform payoff given by Equation 90 with respect to p_c, then the platform's fees which satisfy the first order condition are given as follows:

$$\frac{1}{\pi} \times \frac{\partial \pi}{\partial p_c} = \frac{1}{(p_c + p_s - f)} + \frac{\partial N_{c|q}/\partial p_c}{N_{c|q}} + \frac{\partial N_{s|q}/\partial p_c}{N_{s|q}} = 0 \tag{91}$$

$$\frac{1}{\pi} \times \frac{\partial \pi}{\partial p_c} = \frac{1}{(p_c + p_s - f)} + \frac{-\eta^c_{p_c}}{p_c} + \frac{-\eta^p_{p_c}}{p_c} = 0 \;\; \Rightarrow \;\; p_c + p_s - f = \frac{p_c}{\eta^c_{p_c} + \eta^p_{p_c}} \tag{92}$$

By considering Equation 87 and substituting it in Equation 92, then

$$p_c + p_s - f = \frac{p_c}{(1 + \alpha_s)\eta^c_{p_c}} = \frac{\alpha_s p_c}{(1 + \alpha_s)\eta^p_{p_c}} \tag{93}$$

Similarly, by deriving the platform payoff given by Equation 90 with respect to p_s, then the platform's fees which satisfy the first order condition are given as follows:

$$\frac{1}{\pi} \times \frac{\partial \pi}{\partial p_s} = \frac{1}{(p_c + p_s - f)} + \frac{\partial N_{c|q}/\partial p_s}{N_{c|q}} + \frac{\partial N_{s|q}/\partial p_s}{N_{s|q}} = 0 \tag{94}$$

$$\frac{1}{\pi} \times \frac{\partial \pi}{\partial p_s} = \frac{1}{(p_c + p_s - f)} + \frac{-\eta^c_{p_s}}{p_s} + \frac{-\eta^p_{p_s}}{p_s} = 0 \;\; \Rightarrow \;\; p_c + p_s - f = \frac{p_s}{\eta^c_{p_s} + \eta^p_{p_s}} \tag{95}$$

By considering Equation 89 and substituting it in Equation 95, then

$$p_c + p_s - f = \frac{p_s}{(1 + \alpha_c)\eta^p_{p_s}} = \frac{\alpha_c p_s}{(1 + \alpha_c)\eta^c_{p_s}} \tag{96}$$

By combining Equation 93 and Equation 96, then

$$p_c + p_s - f = \frac{p_c}{(1 + \alpha_s)\eta^c_{p_c}} = \frac{\alpha_s p_c}{(1 + \alpha_s)\eta^p_{p_c}} = \frac{p_s}{(1 + \alpha_c)\eta^p_{p_s}} = \frac{\alpha_c p_s}{(1 + \alpha_c)\eta^c_{p_s}} \tag{97}$$

A.1.2 Proof of Proposition 1.2

By consider Equation 90 again, then the platform payoff can be rewritten as a function of total price $p_t = p_c + p_s$ as follows.

$$\log \pi = \log \gamma + \log(p_t - f) + \log N_{c|q} + \log N_{s|q} \tag{98}$$

By deriving the platform payoff given by Equation 98 with respect to the total price p_t, and considering Equation 10, then the platform's total price which satisfies the first order condition are given as follows:

$$\frac{1}{\pi} \times \frac{\partial \pi}{\partial p_t} = \frac{1}{(p_t - f)} + \frac{\partial N_{c|q}/\partial p_t}{N_{c|q}} + \frac{\partial N_{s|q}/\partial p_t}{N_{s|q}} = 0 \tag{99}$$

$$p_t - f = \frac{-N_{c|q} N_{s|q}}{N_{s|q}(\partial N_{c|q}/\partial p_t) + (\partial N_{s|q}/\partial p_t) N_{c|q}} \qquad = 0 \;\; \Rightarrow \;\; p_t - f = \frac{p_t}{\eta^T_{p_t}} \tag{100}$$

By considering Equation 97, the elasticities of the demands for market sides are given as follows:

$$p_t - f = \frac{p_t}{\eta^T_{p_t}} = \frac{p_c}{(1+\alpha_s)\eta^c_{p_c}} = \frac{\alpha_s p_c}{(1+\alpha_s)\eta^p_{p_c}} = \frac{p_s}{(1+\alpha_c)\eta^p_{p_s}} = \frac{\alpha_c p_s}{(1+\alpha_c)\eta^c_{p_s}} \tag{101}$$

$$\eta^c_{p_c} = \frac{1}{1+\alpha_s} \times \frac{p_c}{p_t} \times \eta^T_{p_t} \tag{102}$$

$$\eta^p_{p_s} = \frac{1}{1+\alpha_c} \times \frac{p_s}{p_t} \times \eta^T_{p_t} \tag{103}$$

$$\eta^p_{p_c} = \frac{\alpha_s}{1+\alpha_s} \times \frac{p_c}{p_t} \times \eta^T_{p_t} \tag{104}$$

$$\eta^c_{p_s} = \frac{\alpha_c}{1+\alpha_c} \times \frac{p_s}{p_t} \times \eta^T_{p_t} \tag{105}$$

By finding the summation of all elasticities, then

$$\eta^c_{p_s} + \eta^p_{p_s} + \eta^p_{p_s} + \eta^p_{p_c} = \eta^T_{p_t} \tag{106}$$

A.2 Extraction of Assumption 1

By substituting separately the p_s and p_c given in Equations 12 and 13 in the platform payoff given by Equation 11, the value of the platform payoff π^* is as given in Equation 107 or by Equation 108. i.e. substitute Equation 12 in Equation 11 to obtain Equation 108, and thereafter substitute Equation 13 in Equation 11 to obtain Equation 107. Since the platform should receive positive payoff to work sufficiently, π^* is positive.

$$\pi^* = \frac{-\gamma N_{s|q} N_{c|q} (\alpha_s \alpha_c - 1) p_s}{B_s(1 + \alpha_c)} \geq 0 \tag{107}$$

$$\pi^* = \frac{-\gamma N_{s|q} N_{c|q} (\alpha_s \alpha_c - 1) p_c}{B_c(1 + \alpha_s)} \geq 0 \tag{108}$$

For externalities $(\alpha_c \alpha_s - 1) > 0$, optimal fees are negatives, $p_c, p_s \leq 0$, which implies that the platform receives negative payoff. Thus we assume $(\alpha_c \alpha_s - 1) < 0$. For $(\alpha_c \alpha_s - 1) < 0$, optimal fees are positives, $p_c, p_s \geq 0$. Thus, we can conclude that platform never charges market sides with negative fees.

A.3 Proof of Proposition 2

As mentioned earlier in Assumption 1, p_s and p_c are positives and by considering Equation 12 and Equation 13, we can rewrite the optimal fees p_s and p_c that the platform chooses as given in Equations 109 and 110 respectively.

$$p_s = \frac{B_s(\alpha_c + 1)(f - p_c)}{B_s(\alpha_c + 1) + (\alpha_c \alpha_s - 1)} \geq 0 \tag{109}$$

$$p_c = \frac{B_c(\alpha_s + 1)(f - p_s)}{B_c(\alpha_s + 1) + (\alpha_c \alpha_s - 1)} \geq 0 \tag{110}$$

By solving Equation 109 and Equation 110 simultaneously we have:

$$p_s = \frac{B_s(\alpha_c + 1) f}{B_s(\alpha_c + 1) + B_c(\alpha_s + 1) + (\alpha_c \alpha_s - 1)} \geq 0 \tag{111}$$

$$p_c = \frac{B_c(\alpha_s + 1)f}{B_c(\alpha_s + 1) + B_s(\alpha_c + 1) + (\alpha_c\alpha_s - 1)} \geq 0 \tag{112}$$

By studying signs of Equations 110, 112, 109, and 111, we derive conditions that make p_c, $p_s \geq 0$ as follows:

i. $p_c \leq f$ for all externalities $\alpha_c\alpha_s - 1 \in (-B_s(\alpha_c + 1), 0]$

ii. $p_c \geq f$ for all externalities $\alpha_c\alpha_s - 1 \in (-(B_s(\alpha_c + 1) + B_c(\alpha_s + 1)), -B_s(\alpha_c + 1))$

iii. $p_s \leq f$ for all externalities $\alpha_c\alpha_s - 1 \in (-B_c(\alpha_s + 1), 0]$

iv. $p_s \geq f$ for all externalities $\alpha_c\alpha_s - 1 \in (-(B_s(\alpha_c + 1) + B_c(\alpha_s + 1)), -B_c(\alpha_s + 1))$

A.4 Proof of Lemma 1

To study the behavior of the platform over externalities values, (i.e) the optimal fees p_c and p_s that the platform chooses, let us assume that α_c is a constant and α_s is a variable. By deriving Equation 111 and Equation 112 with respect to α_s over the interval $\alpha_c\alpha_s - 1 \in [-(B_s(\alpha_c + 1) + B_c(\alpha_s + 1)), 0]$, we have the first derivative of p_s and p_c with respect to α_s as given in Equations 113 and 114 respectively.

$$\frac{\partial p_s}{\partial \alpha_s} = \frac{-B_s(\alpha_c + 1)(B_c + \alpha_c)f}{[B_s(\alpha_c + 1) + B_c(\alpha_s + 1) + (\alpha_c\alpha_s - 1)]^2} \tag{113}$$

$$\frac{\partial p_c}{\partial \alpha_s} = \frac{B_c(\alpha_c + 1)(B_s - 1)f}{[B_s(\alpha_c + 1) + B_c(\alpha_s + 1) + (\alpha_c\alpha_s - 1)]^2} \tag{114}$$

Similarly, let us assume that α_c is a variable and α_s is a constant. By deriving Equation 112 and Equation 111 with respect to α_c over the interval $\alpha_c\alpha_s - 1 \in [-(B_s(\alpha_c + 1) + B_c(\alpha_s + 1)), 0]$, we have the first derivative of p_s and p_c with respect to α_c as given in Equations 115 and 116 respectively.

$$\frac{\partial p_s}{\partial \alpha_c} = \frac{B_s(\alpha_s + 1)(B_c - 1)f}{[B_s(\alpha_c + 1) + B_c(\alpha_s + 1) + (\alpha_c\alpha_s - 1)]^2} \tag{115}$$

$$\frac{\partial p_c}{\partial \alpha_c} = \frac{-B_c(\alpha_s+1)(B_s+\alpha_s)f}{[B_s(\alpha_c+1)+B_c(\alpha_s+1)+(\alpha_c\alpha_s-1)]^2} \tag{116}$$

Since the first derivative of p_c and p_s given in Equations 116 and 113 are negatives, then 1) p_c and p_s are non-increasing over externalities values $\alpha_c\alpha_s - 1 \in [-(B_s(\alpha_c+1)+B_c(\alpha_s+1)), 0]$; and 2) signs of Equations 114 and 115 are negative too. Thus, we assume that B_s, $B_c \leq 1$. The platform payoff around critical points as follows:

$$\lim_{\alpha_c\alpha_s-1\to 0}\left(\frac{B_s(\alpha_c+1)f+B_c(\alpha_s+1)f}{B_s(\alpha_c+1)+B_c(\alpha_s+1)+(\alpha_c\alpha_s-1)} - f\right)N_{c|q}N_{s|q} = 0 \tag{117}$$

$$\lim_{\alpha_c\alpha_s-1\to -(B_s(\alpha_c+1))}\left(\frac{B_s(\alpha_c+1)f+B_c(\alpha_s+1)f}{B_s(\alpha_c+1)+B_c(\alpha_s+1)+(\alpha_c\alpha_s-1)} - f\right)N_{c|q}N_{s|q} = \frac{B_s(\alpha_c+1)f}{B_c(\alpha_s+1)}N_{c|q}N_{s|q} \tag{118}$$

$$\lim_{\alpha_c\alpha_s-1\to -(B_c(\alpha_s+1))}\left(\frac{B_s(\alpha_c+1)f+B_c(\alpha_s+1)f}{B_s(\alpha_c+1)+B_c(\alpha_s+1)+(\alpha_c\alpha_s-1)} - f\right)N_{c|q}N_{s|q} = \frac{B_c(\alpha_s+1)f}{B_s(\alpha_c+1)}N_{c|q}N_{s|q} \tag{119}$$

$$\lim_{\alpha_c\alpha_s-1\to -(B_s(\alpha_c+1)+B_c(\alpha_s+1))}\left(\frac{B_s(\alpha_c+1)f+B_c(\alpha_s+1)f}{B_s(\alpha_c+1)+B_c(\alpha_s+1)+(\alpha_c\alpha_s-1)} - f\right)N_{c|q}N_{s|q} = \infty \tag{120}$$

Since p_c and p_s are non-increasing as derived from their derivatives, and since values of platform payoff around are critical points are increasing as derived from their limits. We state that the platform receives positive payoff over $\alpha_c\alpha_s - 1 \in [-(B_s(\alpha_c+1)+B_c(\alpha_s+1)), 0]$. This is the proof of point 1.

By exploring the sign of Equation 112 and Equation 111, p_c and p_s are negatives when $\alpha_c\alpha_s - 1 < -(B_s(\alpha_c+1)+B_c(\alpha_s+1))$. Which means the platform receives negative payoff in this range of externalities. Thus the platform will not enter the market if externalities between market sides falls in this range. This is the proof of point 2.

A.5 Proof of Lemma 2

According to the results concluded in Proposition 2, when $-B_s(\alpha_c + 1) > -B_c(\alpha_s + 1)$, then we have $p_c \geq f$ and $p_s \leq f$ for all externalities between $-B_s(\alpha_c + 1)$ and $-B_c(\alpha_s + 1)$. Since $p_s \leq f$ and $p_c \geq f$, we can conclude that the platform subsidizes the providers side and make profit from the consumer side. Thus, the providers side is subsidized if $-B_s(\alpha_c + 1) > -B_c(\alpha_s + 1)$. Similary, when $-B_s(\alpha_c + 1) < -B_c(\alpha_s + 1)$, then we have $p_s \geq f$ and $p_c \leq f$ for all externalities between $-B_s(\alpha_c + 1)$ and $-B_c(\alpha_s + 1)$. (i.e) consumer sides is subsided.

By considering Equation 87 and Equation 89, the externalities of market sides and slopes of their curves with respect to transaction fees are given by their elasticities as follows.

$$\alpha_s = \frac{\eta^p_{p_c}}{\eta^c_{p_c}} \tag{121}$$

$$\alpha_c = \frac{\eta^c_{p_s}}{\eta^p_{p_s}} \tag{122}$$

By considering Equation 86 and Equation 88, then

$$\beta_c = -(\alpha_c\alpha_s - 1)\eta^c_{p_c} \tag{123}$$

$$\beta_s = -(\alpha_c\alpha_s - 1)\eta^p_{p_s} \tag{124}$$

Thus,

$$-B_c(\alpha_s + 1) < -B_s(\alpha_c + 1) \Rightarrow \eta^c_{p_c} - \eta^c_{p_s} < \eta^p_{p_s} - \eta^p_{p_c}$$

A.6 Proof of Theorem 1

After many transactions over the platform, data consumers and providers negotiate each other to interact directly without the platform. One of them bears paying a per transaction incentive p_i to convince the other to interact directly. Let us assume that consumers pay p_i for providers (we will get the same result if we assume that providers pay p_i for consumers). Consumers and providers incur a transaction costs $f_c \geq 0$, $f_s \geq 0$ respectively if they agree the direct interaction. Consumers

and providers agree to interact over the platform if they receive more utilities than utilities received from the interacting directly. Thus,

$$v_i(n|q) - n(p + p_c) \geq v_i(n'|q) - n'(p + p_i + f_c) \text{ for the consumer } i$$

$$\gamma N_{c|q}(p - p_s) \geq \gamma N_{c|q}(p + p_i - f_s) \text{ for each provider } j$$

The platform wants to keep both sides interacting through it, and maximizes his utility at the same time. Thus, the platform choose optimal fees p_c^*, p_s^* as follows.

$$p_s^* = f_s - p_i \tag{125}$$

$$p_c^* = \frac{v_i(n|q) - v_i(n'|q) - np + n'(p + p_i + f_c)}{n} \tag{126}$$

A.7 Proof of Corollary 1

By substituting the value functions $v_i(n|q)$ and $v(n'|q)$ given by Equation 1 in Equation 126, we have:

$$p_c^* = \frac{alog(n/n')}{n} - \frac{(n - n')p}{n} + \frac{n'(p_i + f_c)}{n} \tag{127}$$

The platform receives payoff π^* from n transaction performed by consumer i as follows:

$$\pi^* = (p_c^* + p_s^* - f)n \tag{128}$$

By substituting Equations 125 and 127 in Equation 128, we obtain:

$$\pi^* = a\log(n/n') - np + n'(p + p_i + f_c) + n(f_s - p_i) - nf \tag{129}$$

By deriving π^* with respect to n and equaling the derivative to zero ($\frac{\partial \pi^*}{\partial n} = 0$), we have the optimal n^* given as:

$$n^* = \frac{a}{p + p_i + f - f_s} \tag{130}$$

By checking the sign of the second derivative, we find it negative. Thus π^* has maximum value at the optimal n^*. By substituting the optimal n^* in Equation 129 and recalling f = f_s + f_c, we have

the maximum payoff for the platform as given in Equation 131.

$$\pi^* = a(\log(n/n') + \frac{n'}{n} - 1) \tag{131}$$

To make the platform is sufficient, the platform has to receive a positive payoff. Thus

$$\log(n/n') > 1 - \tfrac{n'}{n} > 0$$

$$\Rightarrow (n/n') > 1$$

For $N_{c|q}$ consumers requiring n amount of data. We have total platform payoff given as follows:

$$\pi^*_{total} = \sum_{i=1}^{N_{c|q}} a(\log(n_i/n'_i) + \frac{n'_i}{n_i} - 1) \tag{132}$$

n' increases in each time consumers connect and perform transactions over the platform because they know more providers. (i.e) the size of the connections network $N^{s'}$ for consumer increases. As explained earlier, consumers have sufficient range of the data amount ,$[min_n, max_n]$. When n and n' close to max_n, then the total platform payoff as follows:

$$\sum_{i=1}^{N_{c|q}} \lim_{(n_i,n'_i)\to(max_n,max_n)} a(\log(n/n') + \frac{n'}{n} - 1) = 0 \tag{133}$$

Thus the platform can not attract consumers any more once consumers can receive the maximum sufficient data form the direct sale. We will get the same result if we assume that providers pay per transaction incentive p_i for consumers.

A.8 Proof of Lemma 3

Assume max_n is the maximum sufficient amount of data that consumer i requires for accomplish a certain task. For the first time the consumer i requests data from the platform, max_n providers are assigned and introduced for the consumer i. For the second data request, $x\,max_n$ providers are available from the first time. Thus, the platform match and introduce $max_n - x\,max_n$ new providers for the consumer i. For the third data request, $((x\,max_n) + (max_n - x\,max_n))x$ providers are available from the first and second time. Thus, the platform match and introduce

$max_n - ((x\,max_n) + (max_n - x\,max_n))x$ new providers for the consumer i. We define the function $f(.)$ which represents the number of new providers that the consumer i will know from each data request. Starting the index of requests from zero, $f(.)$ is given as follows:

$$f(0) = (max_n)x$$

$$f(1) = (max_n - f(0))x = max_n(1-x)x$$

$$f(2) = (max_n - f(0) - f(1))x = max_n(1-x)^2x$$

$$f(n) = (max_n - \textstyle\sum_{j=1}^{k=n-1} f(j))x = max_n(1-x)^nx$$

Based on Theorem 1, the consumer i stops requesting data from the platform and interacts directly with providers if he receives the maximum sufficient data from the direct sale. The consumer i receives maximum sufficient data (max_n) after performing r requests over the platform as follows:

$$max_n + \Delta = \sum_{j=0}^{r} max_n(1-x)^j x \tag{134}$$

we can rewrite Equation 134 as follows:

$$max_n + \Delta = f(0) + max_n x \sum_{j=1}^{r} (1-x)^j \tag{135}$$

Recall the geometric series:

$$\sum_{j=1}^{n} ak = a + ak + ak^2 + \cdots + ak^{n-1} = a\frac{1-k^n}{1-k} \tag{136}$$

Using 136, we can rewrite Equation 135 as follows:

$$max_n + \Delta = f(0) + max_n x(1-x)\frac{1-(1-x)^r}{x} \tag{137}$$

By solving Equation 135, the number of requests r is given as follows:

$$r = \frac{2(1-x) - \frac{max_n + \Delta}{max_n}}{\log(1-x)} \tag{138}$$

Table A.1 simulates the values of r over different ranges of x and max_n.

Table A.1: Required period to perform r in years, $\lambda = 24$ requests / day

max_n \ x	1×10^{-5}	5×10^{-5}	1×10^{-4}	5×10^{-4}	1×10^{-3}	5×10^{-3}	1×10^{-2}	5×10^{-2}	1×10^{-1}
1×10^4	266.5	53.29	26.6	5.3	2.6	0.53	0.26	0.05	0.025
5×10^3	250.45	50.09	25.04	5.00	2.50	0.5	0.24	0.04	0.024
1×10^3	213.20	42.63	21.31	4.26	2.13	0.42	0.21	0.04	0.020
5×10^2	197.15	39.43	19.71	3.94	1.97	0.39	0.19	0.038	0.018
1×10^2	159.90	31.97	15.98	3.19	1.59	0.31	0.15	0.031	0.015

Bibliography

[1] Blockchain and big data: the match made in heavens. https://towardsdatascience.com/blockchain-and-big-data-the-match-made-in-heavens-337887a0ce73 . Accessed: 2019-01-02.

[2] Dmr amazon statistical report 2018. https://expandedramblings.com/index.php/downloads/dmr-amazon-web-services-report/. Accessed: 2019-01-31.

[3] Facebook. http://investor.fb.com/releasedetail.cfm?ReleaseID=893395. Accessed: 2016-01-25.

[4] Resdac : Research data assistance center. http://www.resdac.org/. Accessed: 2016-01-25.

[5] Anish Agarwal, Munther Dahleh, Thibaut Horel, and Maryann Rui. Towards data auctions with externalities, 2020.

[6] Anish Agarwal, Munther Dahleh, and Tuhin Sarkar. A marketplace for data: An algorithmic solution. In *Proceedings of the 2019 ACM Conference on Economics and Computation*, EC '19, page 701–726, New York, NY, USA, 2019. Association for Computing Machinery.

[7] Wael AlRahal AlOrabi, Sawsan Abdul Rahman, May El Barachi, and Azzam Mourad. Towards on demand road condition monitoring using mobile phone sensing as a service. *Procedia Computer Science*, 83(Supplement C):345 – 352, 2016. The 7th International Conference on Ambient Systems, Networks and Technologies (ANT 2016) / The 6th International Conference on Sustainable Energy Information Technology (SEIT-2016) / Affiliated Workshops.

[8] Fadi Alzhouri, Anjali Agarwal, Yan Liu, and Ahmed Saleh Bataineh. Dynamic pricing for maximizing cloud revenue: A column generation approach. In *Proceedings of the 18th International Conference on Distributed Computing and Networking*, ICDCN '17, New York, NY, USA, 2017. Association for Computing Machinery.

[9] Amazon. IoT and big data services in Amazon market places. https://aws.amazon.com/marketplace/search?page=1&category=96c2cd16-fe69-4b18-99cc-e016c61e820c. [Online; accessed 19-Nov-2019].

[10] Amazon. Simple monthly calculator. https://calculator.s3.amazonaws.com/index.html. [Online; accessed 19-July-2019].

[11] Hagiu Andrei. Merchant or two-sided platform? *Review of Network Economics*, 6(2):1–19, 2007.

[12] Mark Armstrong. Competition in two-sided markets. *The RAND Journal of Economics*, 37(3):668–691, 2006.

[13] Ahmed Saleh Bataineh, Jamal Bentahar, Omar Abdel Wahab, Rabeb Mizouni, and Gaith Rjoub. A game-based secure trading of big data and iot services: Blockchain as a two-sided market. In *Service-Oriented Computing*, pages 85–100, Cham, 2020. Springer International Publishing.

[14] Ahmed Saleh Bataineh, Jamal Bentahar, Rabeb Mizouni, Omar Abdel Wahab, Gaith Rjoub, and May El Barachi. Cloud computing as a platform for monetizing data services: A two-sided game business model, 2021.

[15] Ahmed Saleh Bataineh, Jamal Bentahar, Omar Abdel Wahab, Rabeb Mizouni, and Gaith Rjoub. Cloud as platform for monetizing complementary data for ai-driven services: A two-sided cooperative game. In *2021 IEEE International Conference on Services Computing (SCC)*, pages 443–449, 2021.

[16] Ahmed Saleh Bataineh, Rabeb Mizouni, Jamal Bentahar, and May El Barachi. Toward monetizing personal data: A two-sided market analysis. *Future Generation Computer Systems*, 111:435–459, 2020.

[17] Ahmed Saleh Bataineh, Rabeb Mizouni, May El Barachi, and Jamal Bentahar. Monetizing personal data: A two-sided market approach. In *The 7th International Conference on Ambient Systems, Networks and Technologies (ANT 2016) / The 6th International Conference on Sustainable Energy Information Technology (SEIT-2016) / Affiliated Workshops, May 23-26, 2016, Madrid, Spain*, pages 472–479, 2016.

[18] France Bélanger and Robert E. Crossler. Privacy in the digital age: A review of information privacy research in information systems. *MIS Q.*, 35(4):1017–1042, December 2011.

[19] Vincent D. Blondel, Adeline Decuyper, and Gautier Krings. A survey of results on mobile phone datasets analysis. *CoRR*, abs/1502.03406, 2015.

[20] J. Burke, D. Estrin, M. Hansen, A. Parker, N. Ramanathan, S. Reddy, and M. B. Srivastava. Participatory sensing. In *In: Workshop on World-Sensor-Web (WSW'06): Mobile Device Centric Sensor Networks and Applications*, pages 117–134, 2006.

[21] Bernard Caillaud and Bruno Jullien. Chicken & egg: Competition among intermediation service providers. *RAND Journal of Economics*, 34(2):309–28, 2003.

[22] Andrew T. Campbell, Shane B. Eisenman, Nicholas D. Lane, Emiliano Miluzzo, and Ronald A. Peterson. People-centric urban sensing. In *Proceedings of the 2Nd Annual International Workshop on Wireless Internet*, WICON '06, New York, NY, USA, 2006. ACM.

[23] Jacob Chakareski. Cost and profit driven cloud-p2p interaction. *Peer-to-Peer Networking and Applications*, 8(2):244–259, Mar 2015.

[24] Ambarish Chandra and Allan Collard-Wexler. Mergers in two-sided markets: An application to the canadian newspaper industry. *Journal of Economics & Management Strategy*, 18(4):1045–1070, 2009.

[25] Jean charles Rochet and Jean Tirole. Two-sided markets: An overview, 2004.

[26] Le Chen, Alan Mislove, and Christo Wilson. An empirical analysis of algorithmic pricing on amazon marketplace. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, pages 1339–1349, Republic and Canton of Geneva, Switzerland, 2016. International World Wide Web Conferences Steering Committee.

[27] Ying-Ju Chen, Yves Zenou, and Junjie Zhou. Competitive pricing strategies in social networks. *The RAND Journal of Economics*, 49(3):672–705, 2018.

[28] R. H. Coase. The problem of social cost. *The Journal of Law & Economics*, 3:1–44, 1960.

[29] Sunny Consolvo, David W. McDonald, Tammy Toscos, Mike Y. Chen, Jon Froehlich, Beverly Harrison, Predrag Klasnja, Anthony LaMarca, Louis LeGrand, Ryan Libby, Ian Smith, and James A. Landay. Activity sensing in the wild: A field trial of ubifit garden. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, pages 1797–1806, New York, NY, USA, 2008. ACM.

[30] Balázs Cs. Csáji, Arnaud Browet, V.A. Traag, Jean-Charles Delvenne, Etienne Huens, Paul Van Dooren, Zbigniew Smoreda, and Vincent D. Blondel. Exploring the mobility of

mobile phone users. *Physica A: Statistical Mechanics and its Applications*, 392(6):1459 – 1473, 2013.

[31] Peter J. Danaher. Optimal pricing of new subscription services: Analysis of a market experiment. *Marketing Science*, 21(2):119–138, 2002.

[32] Yves-Alexandre de Montjoye, César A Hidalgo, Michel Verleysen, and Vincent D Blondel. Unique in the crowd: The privacy bounds of human mobility. *Scientific Reports*, 3, 2013.

[33] J. Ding, R. Yu, Y. Zhang, S. Gjessing, and D. H. K. Tsang. Service provider competition and cooperation in cloud-based software defined wireless networks. *IEEE Communications Magazine*, 53(11):134–140, 2015.

[34] Nicholas Economides and Joacim Tåg. Network neutrality on the internet: A two-sided market analysis. *Information Economics and Policy*, 24(2):91–104, 2012.

[35] Xiaowen Gong, Lingjie Duan, Xu Chen, and Junshan Zhang. When social network effect meets congestion effect in wireless networks: Data usage equilibrium and optimal pricing. *IEEE Journal on Selected Areas in Communications*, 35(2):449–462, 2017.

[36] Marta C. Gonzalez, Cesar A. Hidalgo, and Albert-Laszlo Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, June 2008.

[37] Google. Google clustred data. https://github.com/google/cluster-data. [Online; accessed 19-July-2019].

[38] Albert Greenberg, James Hamilton, David A. Maltz, and Parveen Patel. The cost of a cloud: Research problems in data center networks. *SIGCOMM Comput. Commun. Rev.*, 39(1):68–73, December 2008.

[39] Werner Güth, Rolf Schmittberger, and Bernd Schwarze. An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior and Organization*, 3(4):367 – 388, 1982.

[40] R. Herring, A. Hofleitner, S. Amin, T. Abou Nasr, A. Abdel Khalek, P. Abbeel, and A. Bayen. Using mobile phones to forecast arterial traffic through statistical learning. In *Proceedings of the 89th Annual Meeting of the Transportation Research Board*, Washington D.C., 2010.

[41] Ryan Herring, Aude Hofleitner, Dan Work, Olli-Pekka Tossavainen, and Alexandre M. Bayen. Mobile millennium - participatory traffic estimation using mobile phones. 2009.

[42] Jiejun Hu, Kun Yang, Kezhi Wang, and Kai Zhang. A blockchain-based reward mechanism for mobile crowdsensing. *IEEE Transactions on Computational Social Systems*, 7(1):178–191, 2020.

[43] Shenlong Huangfu, Bin Guo, Zhiwen Yu, and Dongsheng Li. Using the model of markets with intermediaries as an incentive scheme for opportunistic social networks. In *Proceedings of the 2013 IEEE 10th International Conference on Ubiquitous Intelligence & Computing and 2013 IEEE 10th International Conference on Autonomic & Trusted Computing*, UIC-ATC '13, pages 142–149, Washington, DC, USA, 2013. IEEE Computer Society.

[44] Kai-Lung Hui and Ivan P. L. Png. *The economics of privacy*, volume 1, chapter 9, pages 471–497. Elsevier, 2006.

[45] Timothy Hunter, Teodor Moldovan, Matei Zaharia, Samy Merzgui, Justin Ma, Michael J. Franklin, Pieter Abbeel, and Alexandre M. Bayen. Scaling the mobile millennium system in the cloud. In *Proceedings of the 2Nd ACM Symposium on Cloud Computing*, SOCC '11, pages 28:1–28:8, New York, NY, USA, 2011. ACM.

[46] Y. Jiao, P. Wang, S. Feng, and D. Niyato. Profit maximization mechanism and data management for data analytics services. *IEEE Internet of Things Journal*, 5(3):2001–2014, 2018.

[47] Y. Jiao, P. Wang, D. Niyato, and Z. Xiong. Social welfare maximization auction in edge computing resource allocation for mobile blockchain. In *2018 IEEE International Conference on Communications (ICC)*, pages 1–6, May 2018.

[48] A. Jin, W. Song, P. Wang, D. Niyato, and P. Ju. Auction mechanisms toward efficient resource sharing for cloudlets in mobile cloud computing. *IEEE Transactions on Services Computing*, 9(6):895–909, 2016.

[49] B. Jullien. *Two-sided markets and electronic intermediaries: presented at CESifo Economic Studies Conference on Understanding the Digital Economy: Facts and Theory, July 2004*. CESifo working paper series. CES, 2004.

[50] Maha Kadadha, Hadi Otrok, Rabeb Mizouni, Shakti Singh, and Anis Ouali. Sensechain: A blockchain-based crowdsensing framework for multiple requesters and multiple workers. *Future Generation Computer Systems*, 105:650–664, 2020.

[51] Andreas Krause, Eric Horvitz, Aman Kansal, and Feng Zhao. Toward community sensing. In *Proceedings of the 7th International Conference on Information Processing in Sensor Networks*, IPSN '08, pages 481–492, Washington, DC, USA, 2008. IEEE Computer Society.

[52] Kenneth C. Laudon. Markets and privacy. *Commun. ACM*, 39(9):92–104, September 1996.

[53] Juong-Sik Lee and Baik Hoh. Sell your experiences: a market mechanism based incentive for participatory sensing. In *PerCom*, pages 60–68. IEEE Computer Society, 2010.

[54] Xiao-Bai Li and Varghese S. Jacob. Adaptive data reduction for large-scale transaction data. *European Journal of Operational Research*, 188(3):910–924, 2008.

[55] Xiao-Bai Li and Srinivasan Raghunathan. Pricing and disseminating customer data with privacy awareness. *Decis. Support Syst.*, 59:63–73, March 2014.

[56] F. Liang, W. Yu, D. An, Q. Yang, X. Fu, and W. Zhao. A survey on big data market: Pricing, trading and protection. *IEEE Access*, 6:15132–15154, 2018.

[57] Y. Lin and H. Shen. Autotune: game-based adaptive bitrate streaming in p2p-assisted cloud-based vod systems. In *2015 IEEE International Conference on Peer-to-Peer Computing (P2P)*, pages 1–10, 2015.

[58] Z. Liu, N. C. Luong, W. Wang, D. Niyato, P. Wang, Y. Liang, and D. I. Kim. A survey on blockchain: A game theoretical perspective. *IEEE Access*, 7:47615–47643, 2019.

[59] N. C. Luong, D. T. Hoang, P. Wang, D. Niyato, D. I. Kim, and Z. Han. Data collection and wireless communication in internet of things (iot) using economic analysis and pricing models: A survey. *IEEE Communications Surveys Tutorials*, 18(4):2546–2590, 2016.

[60] N. C. Luong, P. Wang, D. Niyato, Y. Wen, and Z. Han. Resource management in cloud networking using economic analysis and pricing models: A survey. *IEEE Communications Surveys Tutorials*, 19(2):954–1001, 2017.

[61] N. C. Luong, Z. Xiong, P. Wang, and D. Niyato. Optimal auction for edge computing resource management in mobile blockchain networks: A deep learning approach. In *2018 IEEE International Conference on Communications (ICC)*, pages 1–6, May 2018.

[62] Lena Mashayekhy, Mahyar Movahed Nejad, and Daniel Grosu. A two-sided market mechanism for trading big data computing commodities. *2014 IEEE International Conference on Big Data (Big Data)*, pages 153–158, 2014.

[63] Emiliano Miluzzo, Nicholas D. Lane, Kristóf Fodor, Ronald Peterson, Hong Lu, Mirco Musolesi, Shane B. Eisenman, Xiao Zheng, and Andrew T. Campbell. Sensing meets mobile social networks: The design, implementation and evaluation of the cenceme application. In *Proceedings of the 6th ACM Conference on Embedded Network Sensor Systems*, SenSys '08, pages 337–350, New York, NY, USA, 2008. ACM.

[64] Song min Kim. How can we make a socially optimal large-scale media platform? analysis of a monopolistic internet media platform using two-sided market theory. *Telecommunications Policy*, 40(9):899 – 918, 2016.

[65] Prashanth Mohan, Venkata N. Padmanabhan, and Ramachandran Ramjee. Nericell: Rich monitoring of road and traffic conditions using mobile smartphones. In *Proceedings of the 6th ACM Conference on Embedded Network Sensor Systems*, SenSys '08, pages 323–336, New York, NY, USA, 2008. ACM.

[66] K. Sridhar Moorthy. Market segmentation, self-selection, and product line design. *Marketing Science*, 3(4):288–307, 1984.

[67] Min Mun, Sasank Reddy, Katie Shilton, Nathan Yau, Jeff Burke, Deborah Estrin, Mark Hansen, Eric Howard, Ruth West, and Péter Boda. Peir, the personal environmental impact report, as a platform for participatory sensing systems research. In *Proceedings of the 7th International Conference on Mobile Systems, Applications, and Services*, MobiSys '09, pages 55–68, New York, NY, USA, 2009. ACM.

[68] Michael Mussa and Sherwin Rosen. Monopoly and product quality. *Journal of Economic Theory*, 18(2):301–317, 1978.

[69] Arvind Narayanan and Vitaly Shmatikov. De-anonymizing social networks. In *Proceedings of the 2009 30th IEEE Symposium on Security and Privacy*, SP '09, pages 173–187, Washington, DC, USA, 2009. IEEE Computer Society.

[70] D. Niyato, D. T. Hoang, N. C. Luong, P. Wang, D. I. Kim, and Z. Han. Smart data pricing models for the internet of things: a bundling strategy approach. *IEEE Network*, 30(2):18–25, 2016.

[71] Ranjan Pal and Pan Hui. Economic models for cloud service markets: Pricing and capacity planning. *Theoretical Computer Science*, 496:113 – 124, 2013.

[72] K. Pantelis and L. Aija. Understanding the value of (big) data. In *2013 IEEE International Conference on Big Data*, pages 38–42, Silicon Valley, CA, USA, 2013. IEEE Computer Society.

[73] Njoroge Paul, Ozdaglar Asuman, Stier-Moses Nicolás E., and Weintraub Gabriel Y. Investment in two-sided markets and the net neutrality debate. *Review of Network Economics*, 12(4):355–402, 2014.

[74] S. Rebai, M. Hadji, and D. Zeghlache. Improving profit through cloud federation. In *2015 12th Annual IEEE Consumer Communications and Networking Conference (CCNC)*, pages 732–739, 2015.

[75] Sasank Reddy, Deborah Estrin, and Mani Srivastava. Recruitment framework for participatory sensing data collections. In *Proceedings of the 8th International Conference on Pervasive Computing*, Pervasive'10, pages 138–155, Berlin, Heidelberg, 2010. Springer-Verlag.

[76] Gaith Rjoub, Omar Abdel Wahab, Jamal Bentahar, and Ahmed Bataineh. A trust and energy-aware double deep reinforcement learning scheduling strategy for federated learning on iot devices. In *Service-Oriented Computing*, pages 319–333, Cham, 2020. Springer International Publishing.

[77] Gaith Rjoub and Jamal Bentahar. Cloud task scheduling based on swarm intelligence and machine learning. In *2017 IEEE 5th International Conference on Future Internet of Things and Cloud (FiCloud)*, pages 272–279, 2017.

[78] Gaith Rjoub, Jamal Bentahar, Omar Abdel Wahab, and Ahmed Saleh Bataineh. Deep and reinforcement learning for automated task scheduling in large-scale cloud computing systems. *Concurrency and Computation: Practice and Experience*, 2020.

[79] Gaith Rjoub, Jamal Bentahar, and Omar Abdel Wahab. Bigtrustscheduling: Trust-aware big data task scheduling approach in cloud computing environments. *Future Gener. Comput. Syst.*, 110:1079–1097, 2020.

[80] Gaith Rjoub, Jamal Bentahar, Omar Abdel Wahab, and Ahmed Bataineh. Deep smart scheduling: A deep learning approach for automated big data scheduling over the cloud. In *7th International Conference on Future Internet of Things and Cloud (FiCloud)*, pages 189–196, 2019.

[81] Gaith Rjoub, Omar Abdel Wahab, Jamal Bentahar, and Ahmed Saleh Bataineh. Improving autonomous vehicles safety in snow weather using federated yolo cnn learning. In Jamal Bentahar, Irfan Awan, Muhammad Younas, and Tor-Morten Grønli, editors, *Mobile Web and Intelligent Information Systems*, pages 121–134, Cham, 2021. Springer International Publishing.

[82] Jean Rochet and Jean Tirole. Platform competition in two-sided markets. *Journal of the European Economic Association*, 1(4):990–1029, 2003.

[83] Jean-Charles Rochet and Jean Tirole. Defining two-sided markets, 2004.

[84] Jean-Charles Rochet and Jean Tirole. Two-sided markets: a progress report. *The RAND Journal of Economics*, 37(3):645–667, 2006.

[85] Harichandan Roy, Murat Kantarcioglu, and Latanya Sweeney. *Practical Differentially Private Modeling of Human Movement Data*, pages 170–178. Springer International Publishing, Cham, 2016.

[86] Marc Rysman. Competition between networks: A study of the market for yellow pages. *Review of Economic Studies*, 71(2):483–512, 2004.

[87] Marc Rysman. Competition between networks: A study of the market for yellow pages. *The Review of Economic Studies*, 71(2):483–512, 2004.

[88] P. Samimi and A. Patel. Review of pricing models for grid amp; cloud computing. In *2011 IEEE Symposium on Computers Informatics*, pages 634–639, 2011.

[89] Yaron Singer and Manas Mittal. Pricing mechanisms for crowdsourcing markets. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pages 1157–1166, Republic and Canton of Geneva, Switzerland, 2013. International World Wide Web Conferences Steering Committee.

[90] Adam Smith. *An Inquiry into the Nature and Causes of the Wealth of Nations*. University of Chicago Press, 1977.

[91] Chaoming Song, Tal Koren, Pu Wang, and Albert-László Barabási. Modelling the scaling properties of human mobility. *Nature Physics*, 6(10):818–823, September 2010.

[92] Latanya Sweeney. k-anonymity: A model for protecting privacy 1. *Ieee Security And Privacy*, 10(5):1–14, 2002.

[93] Steven Tadelis. Two-sided e-commerce marketplaces and the future of retailing. In Emek Basker, editor, *Handbook on the Economics of Retailing and Distribution*, pages 455–474. Edward Elgar Publishing, 2016.

[94] Arvind Thiagarajan, Lenin Ravindranath, Katrina LaCurts, Samuel Madden, Hari Balakrishnan, Sivan Toledo, and Jakob Eriksson. Vtrack: Accurate, energy-aware road traffic delay estimation using mobile phones. In *Proceedings of the 7th ACM Conference on Embedded Networked Sensor Systems*, SenSys '09, pages 85–98, New York, NY, USA, 2009. ACM.

[95] Valerio Di Valerio, Valeria Cardellini, and Francesco Lo Presti. Optimal pricing and service provisioning strategies in cloud systems: A stackelberg game approach. In *IEEE CLOUD*, pages 115–122. IEEE Computer Society, 2013.

[96] Omar Abdel Wahab, Jamal Bentahar, Hadi Otrok, and Azzam Mourad. A stackelberg game for distributed formation of business-driven services communities. *Expert Systems with Applications*, 45:359–372, 2016.

[97] Omar Abdel Wahab, Azzam Mourad, Hadi Otrok, and Tarik Taleb. Federated machine learning: Survey, multi-level classification, desirable criteria and future directions in communication and networking systems. *IEEE Communications Surveys & Tutorials*, 2021.

[98] Jingzhong Wang, Mengru Li, Yunhua He, Hong Li, Ke Xiao, and Chao Wang. A blockchain based privacy-preserving incentive mechanism in crowdsensing applications. *IEEE Access*, 6:17545–17556, 2018.

[99] Z. Xiong, S. Feng, W. Wang, D. Niyato, P. Wang, and Z. Han. Cloud/fog computing resource management and pricing for blockchain networks. *IEEE Internet of Things Journal*, 6(3):4585–4600, June 2019.

[100] Zehui Xiong, Shaohan Feng, Dusit Niyato, Ping Wang, and Zhu Han. Optimal pricing-based edge computing resource management in mobile blockchain. In *2018 IEEE International Conference on Communications (ICC)*, pages 1–6. IEEE, 2018.

[101] Zehui Xiong, Shaohan Feng, Dusit Niyato, Ping Wang, and Yang Zhang. Competition and cooperation analysis for data sponsored market: A network effects model. *CoRR*, abs/1711.01054, 2017.

[102] C. Xu, K. Wang, P. Li, S. Guo, J. Luo, B. Ye, and M. Guo. Making big data open in edges: A resource-efficient blockchain-based approach. *IEEE Transactions on Parallel and Distributed Systems*, 30(4):870–882, April 2019.

[103] H. Xu and B. Li. A general and practical datacenter selection framework for cloud services. In *2012 IEEE Fifth International Conference on Cloud Computing*, pages 9–16, 2012.

[104] Dejun Yang, Guoliang Xue, Xi Fang, and Jian Tang. Crowdsourcing to smartphones: Incentive mechanism design for mobile phone sensing. In *Proceedings of the 18th Annual International Conference on Mobile Computing and Networking*, Mobicom '12, pages 173–184, New York, NY, USA, 2012. ACM.

[105] Mengmeng Yang, Tianqing Zhu, Kaitai Liang, Wanlei Zhou, and Robert H Deng. A blockchain-based location privacy-preserving crowdsensing system. *Future Generation Computer Systems*, 94:408–418, 2019.

[106] Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. Federated machine learning: Concept and applications. *ACM Trans. Intell. Syst. Technol.*, 10(2), January 2019.

[107] R. Yang, F. R. Yu, P. Si, Z. Yang, and Y. Zhang. Integrated blockchain and edge computing systems: A survey, some research issues and challenges. *IEEE Communications Surveys Tutorials*, 21(2):1508–1532, Secondquarter 2019.

[108] Mengyuan Zhang, Lei Yang, Xiaowen Gong, Shibo He, and Junshan Zhang. Wireless service pricing competition under network effect, congestion effect, and bounded rationality. *IEEE Transactions on Vehicular Technology*, 67(8):7497–7507, 2018.

[109] N. Zhang and H. Hämmäinen. Cost efficiency of sdn in lte-based mobile networks: Case finland. In *2015 International Conference and Workshops on Networked Systems (NetSys)*, pages 1–5, 2015.

[110] Y. Zhang, Z. Xiong, D. Niyato, P. Wang, H. V. Poor, and D. I. Kim. A game-theoretic analysis for complementary and substitutable IoT services delivery with externalities. *IEEE Transactions on Communications*, 68(1):615–629, 2020.

[111] Yang Zhang, Zehui Xiong, Dusit Niyato, Ping Wang, and Jiangming Jin. Joint optimization of information trading in internet of things (iot) market with externalities. In *2018 IEEE Wireless Communications and Networking Conference (WCNC)*, pages 1–6, 2018.