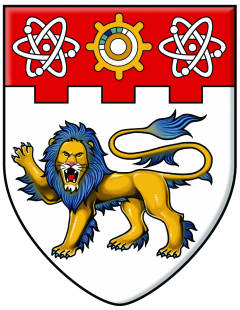


NANYANG TECHNOLOGICAL UNIVERSITY

Monetizing Potential of User Information

Rao Divya Shivkumar

School of Computer Science and Engineering

A thesis submitted to the Nanyang Technological University
in partial fulfilment of the requirement for the degree of
Doctor of Philosophy

2017

Abstract

Information is being shared at an unprecedented rate today and has become the most valuable resource for any organization to help them make better profits. But the lack of a pricing structure or framework affects the users because they don't realize the true value of their information and thus lose out on the monetizing potential of their data.

The research problem before us is, if information is to be treated as a tradeable commodity in the market, how can one put a price tag or value to it? One of the easiest way to communicate value is to put it in terms of price. There is need for models to make users aware of the revenue-generation capacity of their information. But these models also need to be fair and transparent so that the buyers would also be willing to participate in the information market.

In this thesis, we have proposed techniques that borrow from different disciplines to develop pricing models aimed at pricing user information. The idea behind calculating the value of the actual information is based on Shannon's information theory that is utilized to calculate the value of the information attribute. This has been adapted into models from a privacy conscious user's scenario where we supplement it by testing the utility of the user's information using statistical comparison techniques as well as using techniques from finance. For the scenario of a buyer we have complemented this with the Markov process and on the investment optimization approach. We then tested these approaches on datasets and simulation techniques and surveys which have culminated in realizing the value of a user's information. This shows us the potential for revenue generation for a user and also demonstrates reduced prices from the buyer's point of view. Our exploratory models exhibit the capacity to monetize on an internet user's digital footprint.

Big data is already a major source of income for organizations. It is high time that it becomes a source of income for the users who ultimately are the generators of this big data.

Acknowledgements

Words are not enough to express how thankful I am to all those who have been an integral part of my PhD journey.

First and foremost I would like to thank my supervisor, my guide and my teacher Dr. Ng Wee Keong. His guidance and insight have helped shape this thesis. I am grateful for his mentoring which has allowed me to grow and improve myself as a researcher.

I would have absolutely not reached this stage today if not for the unwavering love and support of my family who have always encouraged me throughout.

I would also like to thank Mr. Loo Kian Hock, the administrative and technical staff at the Data Management and Analytics Graduate Lab for his technical assistance and Ms Len Ah Chan, Ms Chiam Poh Ling and Ms Fiona Low Chui Theng for their help in all administrative matters.

My thanks are also due to my ever supportive seniors, Liu Fang and Vasily Sidorov for all their helpful advice.

I am also thankful to my friends who cheered me up and helped keep my spirits high all the time as I navigated this journey. I cannot imagine how I could have done this without Ronak Bajaj, Rishabh Ranjan, Vipra Guneta, Abhishek Jain and Achiranshu Garg.

I would also like to thank the Nanyang Technological University & the School of Computer Science and Engineering for giving me the opportunity to attend one of the best universities in the world.

Lastly, I am thankful to the almighty for blessing me with the strength to write this thesis.

Contents

List of Figures

List of Tables

1

Introduction

With increasing avenues for sharing information available today, massive amount of data is being generated and shared by internet users. In addition to this, is the colossal information generated by users through their interaction with various web services (like Amazon, Spotify etc). And all of this information is constantly being collected and collated every minute today by most organizations, which has led to the belief that big data has now become the oil which keeps the wheels of most organizations running. This big data is generally collected by 'data brokers' who then sell this data to whomsoever is willing to pay the right price. Due to the lack of any specific mechanism to ascertain the value of this information, the contributors and owners of this big data, i.e. the internet users from whom this information is collected from, do not realize the potential of their information and how valuable it is and hence

go largely uncompensated. In this thesis, we explain the background and problems associated with putting a price to data or information and then we put forth our idea of monetizing information by treating it like a tradeable commodity or asset to create different information pricing models for different scenarios that offer a fair bargain to both the consumers and buyers of information.

1.1 Background and Motivations

Users today are getting increasingly open about sharing information in the cyber space. This has been bolstered by the increasing number of avenues for them to do the same. This information can range from personal information, insights, and other types of information like number of queries posted in different websites and blogs etc. Thanks to the proliferation of social media and networking sites today, a massive amount of data is being generated by the second. This information is collated, collected and trawled from various sources to construct a 'user profile' of sorts and used to type cast and categorize people. This is done on a regular basis by data brokers; entities whose sole purpose is to collect user information and sell them to interested parties. Organizations have woken up to the fact that information and not just information products or goods can now be traded in the market place as a commodity. Gartner predicted that 30 percent of businesses will be monetizing their information assets directly by 2016[1]. Economic and pricing models in existence currently look at the consumption of these goods and services by consumers (either individual users, groups of user or organizations) and accordingly formulate price plans to incorporate them. Most organizations seem concerned with product and service development catering to the data that is collected and aggregated by them from a variety of sources. That is how they are monetizing on their information assets. While that is commendable, in our opinion, there doesn't seem to a practical model that would value actual user data in a way to benefit both the sellers (the users providing the data) and the buyers (the consumers who would want to purchase this data for their own use).

[1] http://www.gartner.com/newsroom/id/2299315

1.2 Research Objectives

Big Data is now favored in its usage thanks to the fact that having this kind of information about anyone is almost akin to knowing and predicting a person‘s actions, thoughts and behavior. Though this is an alarming thought, it is common knowledge among data brokers who trade this type of big data and the insights from this big data.

The amount of information generated today is tremendous. According to this article, as of 2012, around 2.5 quintillion of data was created everyday [1]. **Appropriate and moderated access and usage of this user data** can help organizations be of better service to the users by catering to the individual needs of the users. The users (i.e., the generators of information) can benefit not only from the enhanced service but also from the **adequate compensation** that they would receive from the organizations interested in their information. Unfortunately, it has been observed that most of this user generated information (which is often of a personal and private nature) is being blatantly used without the knowledge of the users and even if the users do sign some sort of “terms and conditions” contract, they do not realize how exactly their own information is being used by organizations for their own benefit and profit often at the cost of the users. One of the major advantages faced is price discrimination. Price discrimination occurs when a user’s information is surreptitiously used to price goods and services higher or lower. This creates an asymmetry of profit and benefit between the generators and consumers of information which in our opinion, is highly unfair. In our opinion, with the increasing proliferation of the Internet and social media sites, the issue of pricing information requires a special focus in order to stem the resulting abuse of this information.

Our research objective is to design and propose mechanisms using which, we can create models for users(sellers) for different scenarios and allow them to sell their information to prospective buyers. At the same time, we also need to balance the interests of both the parties and facilitate the proper functioning of an information market.Using these models, users should be able to appropriately value their data or

information At the same time, buyers should be able to purchase this information at a *reasonable price* from the sellers and also be aware of the *quality* of the information that they would be purchasing. The intention is thus to be able to monetize information from the user's point of view and at the same time retaining the relevancy and applicability of the information from the buyer's point of view.

1.3 Research Issues

Today, big data is integral to every business and organization, so much so that data brokers have made a business of trading this big data like any other commodity. In turn, the buyers of this big data make massive profits. The only one who loses out on both, the profits and his privacy is the internet user - the generator and owner of this big data.

Our task is to design balanced pricing mechanisms that incorporate the concerns of the stakeholders in the information market where the user can trade his or her information for a certain price. But information is an 'infinite supply' good since generation of information requires practically no effort, multiple copies of data can be made easily and there is a constant supply of ever growing information. Hence the generation of revenue and the achievement of an equilibrium market price by sale of an infinite supply good is different from that of other finite supply goods.

The building of the pricing models presents before us certain research issues. Of these we have identified and isolated the following:

- Incentivization issues - We need to frame the pricing models to not only alleviate the concerns of the stakeholders in the scenario but also incentivize them to participate in the idea of sharing and selling information.
- Valuation - The idea behind a pricing model is to be able to appropriately valuate the information that can satisfy any apprehensions the parties may have. Valuation of information is a difficult task (as elaborated in the later chapters). The seller and the buyer may have different ideas about the price and value of the information. The issue then is to find an method that would appeal to the

private sensibilities of the stakeholders involved in the transaction.

- Information types - The type of information about and contributed by the users (i.e. the sellers) also plays an important role in the pricing model. The different information types could be divided into: textual, media or numeric. Another way to divide the information types could be: general data (age, sex, education), location (GPS co-ordinates), financial (card or cash transactions, type of purchases). The approach on how this information is clubbed together and presented could also play a major role in the pricing model.
- Empirical/Experimental analysis - Lastly, the reliability and the acceptance of the pricing models must be tested among the actual users to see if they are amenable to the idea of selling their information. This testing will also help us understand any gaps or inconsistencies that may be there in the model.

1.4 Research approach and Methodology

With the multitude of channels available today, information is available and flowing more freely than ever before. Combine this with the advancement in algorithms and programs for complex data analysis, today not just raw data, but information obtained from analyzing this raw data can offer major insights into the life of a normal internet user. The problem that this creates is not just related to the invasion of privacy but there is also the issue that users never benefit monetarily similar to the profit organizations gain from applying this information. A monetary benefit is different from a social benefit which a user experiences while using a service or product.

Most of existing systems highlight the need for privacy. What they fail to address is the question of monetary compensation for users. Data brokers collect user data without the knowledge and often without the consent of the users and sell this data to interested buyers. The buyers benefit from having access to all this information which can help them make better decisions about their business models and earn better and increased profits. Often here it is not about the services but rather about classifying the users into “categories” or “bins” (eg: soccer moms, elderly etc) which

would benefit the organizations.

Since the gathering and selling of information has now evolved into a business, we believe that we should treat user information as a commodity or an asset which can be traded in the internet market by the users themselves.

The idea is to involve users and incentivize them to share their information by rewarding them with monetary compensation when they share their information (Eg: bank balance, habits, criminal history, sexual activity etc). This measure can be successful only if we know the value of the information. At the same time, it is also needed to be fair to the buyers of this information by assigning fair prices to this information. This mechanism of assigning prices should be transparent to give the buyers more confidence about the prices and should also be able to convey the idea of information quality.

We have utilized ideas and concepts from information theory, stochastic modeling techniques and also from the field of data privacy to build our pricing models in an effort to allow users to valuate their information.

In this thesis, we shall first be analyzing the existing concepts and mechanisms in different disciplines that make use of the concept of value and pricing. We shall also be evaluating the issues faced with private data in general. In keeping with the digital age, there do exist pricing schemes for information goods (including softwares) but most of the times these cannot be applied directly to raw data or private data due to issues which we shall explain in Chapter 2.

A mechanism for pricing must satisfy both the buyers and the sellers involved in the transaction. Since most of the data that the buyers are interested in is of a private or personal nature, this brings into picture another aspect to be looked into in the scenario of the information market which is, the privacy budget of the seller which details the amount of distortion a buy is willing to tolerate in user information to protect the privacy. This budget must be balanced with the ultimate utility (quality) of data. We can arrive at a satisfactory conclusion only after a thorough analysis of the existing pricing strategies and mechanisms in use today.

With the vast amount of data being shared and updated (similar to [52]) we propose a mechanism that can design a suitable scenario satisfying both the buyers and sellers of data.

1.5 Thesis Organization

The further chapters of the thesis are organized as follows:

- Chapter 2 presents a detailed look at the existing literature and research on the concept of information pricing. We explore the need for information pricing in-depth and also look at the pricing models available for the traditional goods and also for computing resources. We also accommodate the research done of privacy and user behavior and attitude towards privacy.
- Chapter 3 provides an overview and the rationale for the work done in this thesis and for the path taken to tackle the research goals supplemented with a user case study about their attitude towards information as a sale-able good and also explain the various scenarios for our pricing models.
- Chapter 4 considers the scenario of a privacy insensitive user and compensates him with a pricing model that looks towards the valuation of information for the same.
- Chapter 5 gives the scenario for pricing information with a privacy conscious user using the idea of noise obfuscation and for the compensation from the buyer to be proportionate to the quality of the obfuscated information.
- Chapter 6 also deals with a privacy conscious user but in this chapter we compare the information distortion as a risk adjusted measure to determine the price for the buyer.
- Chapter 7 also describes the scenario of a privacy insensitive user but in this case we have treated the user as an agent in the information market which has a specific demand for the information attribute that the user possesses modeled as a Markov process.
- Chapter 8 describes the information pricing model from the point of view of

the buyer by allowing to help him choose the information to invest in and how much to invest in.

- Finally in Chapter 9, we provide the conclusions of our thesis and the reasoning and the motivation of our research topic of information pricing.

1.6 List of publications

- Rao Divya, and Wee Keong Ng. "How much is your information worth??? A method for revenue generation for your information." Big Data (Big Data), 2015 IEEE International Conference on. IEEE, 2015.
- Rao Divya, and Wee Keong Ng. "How to make money from your information and keep your privacy." Big Data (Big Data), 2015 IEEE International Conference on. IEEE, 2015.
- Rao Divya, and Wee Keong Ng. "A user-centric approach to pricing information" IEEE BigDataService 2016
- Rao Divya, and Wee Keong Ng. "Information pricing for the user in the internet information market" 12th World Congress on Services (SERVICES 2016) 2016
- Rao Divya, and Wee Keong Ng. "A method to price your information asset in the information market" IEEE International Congress on Big Data (BigData Congress 2016)
- Rao Divya, and Wee Keong Ng. "Monetizing the Users Information Asset in Internet Information Market" 5th 2016 IEEE International Congress on Big Data (BigData Congress 2016)
- Rao Divya, and Wee Keong Ng. "Information Pricing - A utility based pricing mechanism" IEEE International Conference on Big Data Intelligence and Computing (Datacom) 2016

2

Related Work

The idea of pricing information is still in its nascent stages. In this chapter we have tried to incorporate the existing and relevant work that have a direct or indirect bearing on our research topic of information pricing. We have covered areas that speak of information as a good, the existing pricing models for the smart grid, cloud services, traditional goods and have also looked at the problem of privacy.

2.1 Introduction

The issue of pricing or valuing information is a complicated one because information or data does not have definite predetermined parameters or values on which one can assess it. It does not have any tangible physical properties and its value or importance differs based on different perspectives [112].

To be able to understand this issue thoroughly, one has to look at different interdisciplinary subjects which can help one delve deeper and be able to analyze and consider the different parameters involved, especially when the nature of information to be priced is of a private and personal nature.

The most neglected and overlooked aspect of information value is the amount of utility that is acquired with just access to the said information. George Stigler, the famous economist, makes a strong point about the importance of information [127]. Though he doesn't value information directly, he associates its importance with empowering buyers to identify sellers and ascertain market prices in order to eliminate uncertainty which in the author's opinion is what drives the market prices.

Based on the vast amount of available literature, we have attempted to group them into different subsections that are based on how the concept of "information" and "pricing" have been handled.

Attaching a price to any good or service is related to the compensation for the use of that particular good or service. [91] states that the price of a good or commodity can be defined as the value of what is the good in question and based on the concept of supply and demand for that commodity. All this can be ascertained because the goods in question all have discernible properties and factors like cost of production, amount of supply and demand, other external costs, etc. This makes it easy to arrive at a particular price.

As [112] rightly points out, unlike other goods and commodities, information does not have any tangible properties making it difficult to put a price or to calculate the price or value for information. Though of course, the value of information could be connected to how it helps in the process of decision making [91].

A data broker in today's day and age is a company that is a silent spectator to all your online transactions and often offline transactions also and tries its level best to gather all sorts of information about you [120] (see Figure 2.1). Organizations have supplemented to their existing knowledge-base with increasingly personal and private information of users, sometimes with and often without the permission or knowledge

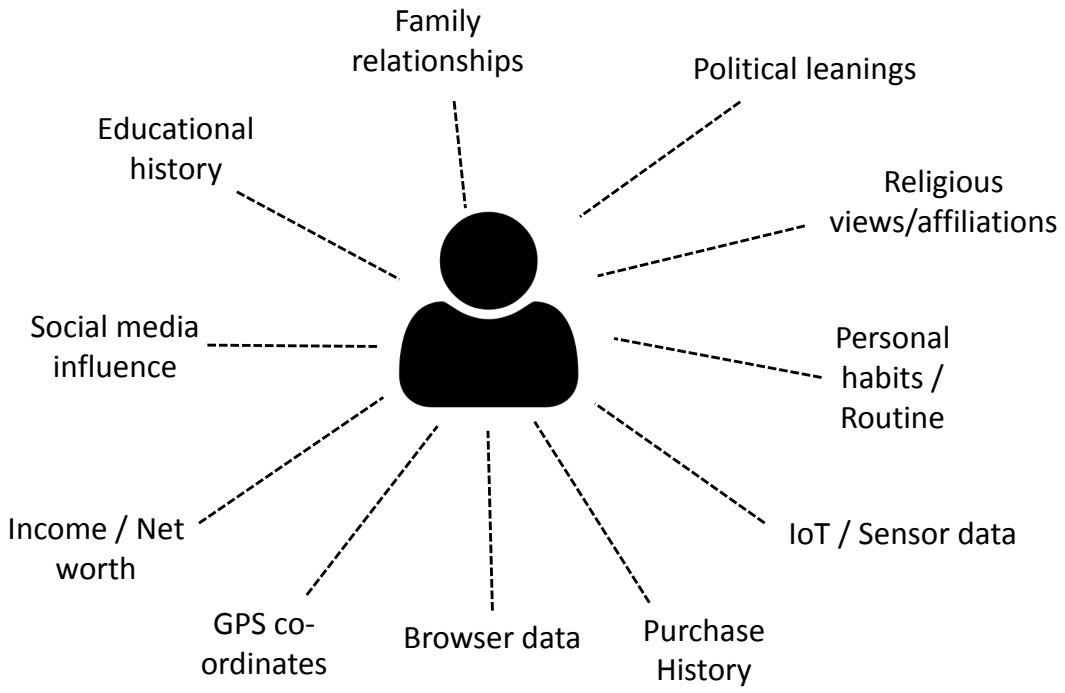


Figure 2.1: Example of the type of information collected by organizations about users

of the users, for the purpose of invasive targeting. This includes information from online surveys, social networks, shopping sites and now a days even the browsing behavior [34, 133]. The organizations then make use of this information to develop "personalized" services for the user who could be charged for them.

Recently, there has been an emergence of a number of tech startups that provide an avenue for users to make their personal data work for them through a number of different applications. Meeco [8] claims to provide a platform using which users can apparently make their information as their currency. Meeco markets itself as a service that can be used by users to manage and organize their information. Users can control access to their information, encrypt it and sync it across all the user devices. But in all this, they make no reference to how this information value translates into price. As of now, they have an iOS app available and for others there is a web version. Another tech startup called 'Datacoup'[1] devised a scheme to pay users for access to their personal data via the user's social media accounts (like Facebook & Twitter) and their credit or debit card feeds [7]. Once they have this information, they strip them off any

[1]http://www.technologyreview.com/news/524621/sell-your-personal-data-for-8-a-month/

personally identifiable information (PII) and sell this information to organizations requiring this information to track trends or to profit from this information. They claim that users can encash their earnings once it reaches $5. Datacoup says that they are lining up purchasers for the user data but currently only Datacoup is the sole buyer. This seems to imply that once Datacoup purchases user data, they can then choose to sell it to whoever and for any price possible. According to the website, their pricing model is based on the individual data attributes shared by each user when the user creates an account with Datacoup. [62] suggests making the internet service providers (ISP), who have the ability to collect user information, give the users the option of either allowing or not allowing their information to be collected. The incentive for the users to allow their information to be collected could be in the form of some free service. But this is not allowing users to monetize on their information.

But the problem with them is that most of them offer services for free [43]. Now this model could hardly be sustainable in the long run and it is believed that ultimately, they may then decide to use their biggest asset, which is a user's personal data, to generate revenue.

In an empirical study conducted by [23], it was found that participants choose to opt for lower price despite having to divulge more personal information. Also, keeping all things equal (price, product, service etc), more privacy friendly measures did not find more favor amongst the participants. It is this dichotomy in user's behavior that has baffled privacy experts for ages and has been a major hurdle in the development of some model to put a price to privacy. This unwillingness to pay for privacy, has made the economics of privacy even more complicated.

2.2 What is Price and What is Value

One way to summarize and define price is to assign it as the value on what is exchanged. Economic value can be related to the amount of benefit that a consumer can obtain from using a particular product or service. [91] says that the price of a commodity depends on the supply of and the demand for that commodity in the market. It

implies that more in demand a commodity, more is the price for that commodity.

One way to describe the concept of pricing is to relate it to the transmission of value that occurs in the exchange of goods that generates economic activity. In [140], the author describes the origins of notion of pricing, explains the labor theory of value and details the different aspects of pricing and what factors affect pricing. The fundamental concepts of a pricing model in regular markets are *rational preference*(meaning that given a choice, a rational person would go for the option that gives him maximum value) and *utility function* (which defines the utility derived from a particular good which in turn determines its value). Other factors affecting pricing are said to be, *asymmetric information* (where one party has more information than the other one; generally in the pricing scenario it means that one party knows something about the product or good in question that the other party does not), *behavioral anomalies* (a part of behavioral finance that looks into phenomenon that cannot be explained by conventional economic principles like rational choice etc), *dynamic pricing* (where the price of a product or good is changed depending on the profile of the customer purchasing the item), and *externalities* (the idea that one agent's reaction could affect the benefit or reward for another agent).

The above works well for regular markets where goods, commodities and services have fixed factors and principles while being traded and there is some precedent that drives the setting of prices. But for information, this falls short because of the inability to pinpoint a set of features that could drive the pricing of information. The take away then is that, for today's data driven society, economic models and theories need to be modified in order to be able to connect them more candidly to data itself because the existing models and theories cannot be applicable when the good in question is information.

2.3 Information - as an economic good

Information has always been recognized as being valuable. It is hard to ignore the immense importance of having access to user information. But it is while putting a

value to information that things get problematic. Information is a peculiar resource with peculiar characteristics. And hence it is unfair to treat information as a traditional good [21]. Information has certain intangible properties that make valuing it complicated. But one area where information can behave like other goods is in the concept of adhering to the principles of supply and demand as is seen today by the varying prices charged by data brokers for different information categories. The value of information has primarily been tied to its effect in decision making [91].

Value of Information (VOI) analysis has been undertaken with the primary goal of calculating the benefit of collecting (additional) information so that decision makers can make an informed decision [146]. VOI illustrates the difference between the'expected' utility of an action with and without the benefit of additional information. It can be said to be the "opportunity loss of taking action under uncertainty". But availing perfect information is a rare possibility and hence VOI considers the expected utility gained under'imperfect information' i.e., action taken in light of the availability of some sample information. But VOI models can get pretty complicated and even for a simple VOI analysis one needs to model the various available actions, previous knowledge about uncertain inputs and the resultant consequences of the actions when decisions are made with these uncertain inputs. A study [59] focused on the impact of decisions on the value of information that considered how the elimination or reduction of uncertainty can improve the probability of outcomes. [66] surveys recently published articles and papers that utilize the concepts of VOI and character izes these articles to understand the existing uses, needs and look at future directions. The authors state that VOI has the ability to manage risk by allowing the decision makes to focus resources on monitoring parameters whose values would cause the most affect the actual results of alternative strategies to secure the system.

For an organization, information has immense importance. And it is upto the CEOs and CFOs to recognize the importance and to take due action to value information. The author [58] states that identification, valuation and ongoing maintainence of the information should be an ongoing process. To identify information value, one can asses the value of an asset to the organization as the difference between the value of

the organisation with the asset, and without.

Information valuation is an intrinsic part of Information Lifecycle Management (ILM) [64]. ILM aims to understand the value of information at different stages in order to enhance the utilization of system resources. Various models have been developed based on different parameters to enforce maximum utilization of the resources at hand. For example, one model is based on the concepts of commodity value and exchange value in economics. This model concerns itself with three aspects - the actual information, information demand (the number of accesses) and the information supply (the competence of the system in providing the required information). Another approach seen is the "usage over time" approach for information valuation in its lifecycle management. In [35], the author incorporates the concept of the ever changing nature of information with time in his model that can help to compare two or more files.

According to T.Nagle in [103] relying solely on economic theory and economic models for pricing decisions is an immature way to look at the pricing problem. Economic theory and the models described are not practical or realistic enough for pricing decisions especially where economics of information is concerned. Though of course in this case the author is referring to the economic value of information with regards an organization's pricing and business decisions.

But apparently, not any kind of information is valuable. For information to have certain value and to be of use, it is essential that it be in the correct format and be corroborated(check for the quality of data) for its genuineness. One study [29] highlighted the problems associated with maintaining data quality and the need for proper valuation of data. They call on the need for'Web observatories' - conglomerate of governmental, research and corporate organizations, to provide some sort of indices in order to evaluate the data quality and provide some sort of a universal compliance methodology. In our opinion, their method of valuing pure 'data - data' exchanges depends entirely on the owners of the data and does not really display the concept of fairness.

Today with the alarming number of security breaches, it is of great importance to look at the economic aspects of information security [17]. When a security breach occurs, the economic losses suffered are far reaching. As explained in [2], it is not just data loss that happens when a data breach occurs in an organization. The far reaching effects of a data breach can last for a long time causing loss of reputation which in turn can cause a loss in the business. Despite increased security measures, it is getting harder to control and prevent data breaches.

A common method to avail information goods and services to all, is to host them on online servers and data bases and associate suitable pricing for these information goods and services [61]. Generally, either a search based or a subscription fee based strategy is used. According to the authors, the strategy must vary keeping in mind the type of users, their information needs, the improvement in technology and so on. But despite this varying of strategy, the revenue in this case is earned only by the one's hosting this information and not the actual owners and contributors of this information.

Network externalities have a considerable effect on the pricing of goods [123]. Network externality in the economic sense, is the change in the benefit that an economic agent would obtain from a good if the number of other participating agents using or consuming that particular kind of good would change. For tangible goods, it is quite simplistic to anticipate the externality involved but with information goods, it becomes quite complicated to calculate the effects of the same on the pricing. It is because of this network externality that firms initially have an introductory or a promotional pricing of the information goods initially in a bid to attract a large user base. Now at times this initial price could be lower than the marginal cost to the firm. But this strategy is not permanent and changes depending on the firm and the information good or product being sold. Often when upgrades are involved, the firms would be profitable if they charge a higher price initially and a lower price for the upgrades.

[138] speaks about pricing in the face of the problems of piracy and the presence

of heterogeneous customers. According to the authors, the rising marginal propensity of consumption of information goods leads to the rise of piracy of these information goods. The marginal propensity to consume is the cumulative proportional rise that occurs in the amount that a consumer would spend on the goods and services instead of saving. They have found that piracy quickens sales times and raises the welfare associated in fixed capacity markets but behaves in a diametrically opposite fashion in growing markets. This paper tries to present a model in which the marginal propensity to pirate can rise in sync with the number of existing current users of a particular information good keeping the heterogeneity of consumers persistent over a time period. The problem is formalized as a profit maximization problem for the legitimate seller who is the one who controls the market sales by controlling the price of the information goods.

[131] acknowledges that the true value of customer information for an organization is related to the ability of these organizations to identify individual customers and then leverage this knowledge to personalize prices to these customers. This paper studies two types of hypothetical settings; one where the sale of customer information is forbidden called as the confidential regime and the other setting where one firm or organization can compile and the sell customer information for the express purpose of implementing price discrimination called as the disclosure regime. The driving force behind this information collation is that a customer's decision to purchase or not to purchase an item (called as a customer's purchasing decision) at one firm is valuable for another firm because each customer has private heterogeneous demands for goods and each customer's valuation for two types of goods end up being positively correlated.

2.4 Existing schemes of pricing computing resources

An optimal pricing policy that balances different factors is often complex and finding that is not the easiest task.

With the coming of the digital age, there has been a growth in the literature that looks at pricing of digital products and Software-as-a-service (SaaS). And traditional pricing concepts cannot be applied directly to software products and services as mentioned in [76]. Similarly, the concepts and parameters mentioned by the author for digital goods like price discrimination (where depending on customer information the items are priced differently), price bundling (certain types of items are sold together or some items are sold together with a discount), cost based pricing (where a certain sum is added to the cost price of the item) and dynamic pricing strategies cannot directly be applied to information or data as a product or good.

A study by [119] critiques the pricing architecture across computer networks which utilizes the idea of utilizing marginal congestion costs for a usage based pricing method to aid in congestion related issues. The authors in this study here mention that these marginal cost based designing techniques (marginal cost is the change that arises in the total cost of an item when the quantity of items produced increases by one unit) may not generate satisfactory revenue and may not aid in eliminating the congestion costs that arise in a computer network. Their proposed model of edge based pricing, though applicable for a "network - based" scenario, does not find application for information as a product to be priced.

The Stackelberg game [137], which forms a part of game theory, is a game in which the leader moves first followed by the other players in the field. Based on this, there is a paper [96] that discusses strategies for information services in the internet of things. In their model, they have three participants, namely the consumer (about whom information is collected using various sensors), the information providers (who provide the consumers with different information services) and the intermediary (who is there to connect the consumers and information providers). Depending on the value of the sensors and the value of the information collected, the information providers use either a bundling strategy or a component strategy. The authors however, fail to provide a clear definition about the consumers and what type of information services they seek and how are the consumers compensated for their information. Also they have not provided any scope for discussion about the privacy of the information being

shared. This is a good example of how user information is leveraged to determine prices for information services or used to run the information services.

A recent attempt (in 2013) has been to study the link between the amount of information collected by advertisers and how effective and profitable it is for them [51]. The authors here have attempted to characterize the revenue generated through advertising as a function of the information collected from the users that is invariably used for online advertising. Their model has three parameters - the users, the publishers and the aggregators. According to their analysis, they conclude that there is a certain bias for some aggregators to collect a certain type of information from certain types of users. Also the information from the users is collected by the aggregators without any concern for the privacy of the information collected. This is why depending on a third party or "middle-man" is a highly risky scenario.

Of late, the market for cloud storage has been increasing rapidly ever since its inception [3] [4]. Most of the firms offering cloud storage offer some amount of free space (around 5 - 10 GB) and after that adopt a competitive pricing scheme [5] [109]. Most of these firms offer something called as a bundling scheme where purchasing two or more items purchased together is cheaper than if they were bought separately. With cloud storage, though privacy and unauthorized access and usage of user information is forbidden, there is always a possibility that some type of user data analysis could be performed by the organizations on all this stored information, though this possibility is never explicitly mentioned [104]. Along the same lines, we have another study [101] that makes a comparative study of the pricing schemes of the various Information-as-a-Service (IaaS) providers. They reported different models and schemes used by different providers. Some of the models mentioned are, the linear model (where the price increases as the amount of the resource (computing resources like CPU, RAM and hard disk space) being utilized increases), the variation of the linear model (where the price is reduced if the user opts for a long term plan of the resource) and finally the step model (where the price reduces as more of the resource is consumed). All of these models have fixed features like CPU usage, hard disk space etc to base the pricing model on.

The pricing scheme of cloud resources has attracted various researchers to study the different schemes and devise an optimized scheme for users and cloud providers. One such work, [87] studies the pricing scheme of Amazon and uses a game theory approach in order to analyze the different approaches used by users to submit their jobs on the cloud. This study focusses on the concept of interruptible services or "spot instances" which are charged based on usage. So this work focuses on the dilemma of how the user should submit the jobs to the cloud resource to gain the maximum benefits from the same at a desirable price.

A common and candid approach by internet based companies offering information as a service and looking to promote their latest good or service, is to offer it free of cost to attract a large user base. The authors of [84] opine that, content providers should strike to achieve a balance between their regular patrons and non - regular patrons by potentially offering the regular ones with a more comprehensive 'free and fee' package and their not so regular users with a simplistic free package of information services. This type of scheme does not consider the value that a user brings in with his information and thus does not look towards compensating the user for the same.

Social networks have a wealth of data available constantly. A study, [32] looks at the pricing of this data from a monopolist's point of view. According to this study, one way to price could be a variation of price discrimination where they can provide a discount depending on how much a user is able to influence other users to purchase a product. Another method could be to offer a single price to everyone. Now this price must be low and optimal enough to attract enough users to enable the firm to make some profit. What has been felt for a long time but has been gaining voice only recently, calls for the development of some sort of a business model to appropriately price high quality information and data [126].

Pricing of resources on the cloud has started attracting attention of various researchers. One study, [145], makes an attempt at understanding the impact of the existing practices on the economical aspect of cloud pricing. Apart from developing a model that looks at the existing resource requirements of different applications, the

researchers also look at potential new research ideas and techniques to improve the revenue of the cloud service providers using methods like throttling and providing performance guarantees. Rather than adopting a similar pricing for everyone as is the usual method, pricing based on resource throttling provides a higher revenues for the cloud provider.

[33] describes a pricing service for grid computing environment by expressing the pricing scheme as an XML document that can be linked to service level arguments. Their pricing scheme dependent on four parameters - quantity, time, quality and user looks towards basing the pricing on the features of grid computing.

The time dependent pricing scheme expressed in [57] tries to address the problem of an ISP's (Internet Service Provider) capacity when usage-based pricing causes the ISPs to swamp the capacity for the peak demand. The author does this by considering when and how much a user consumes data. The paper proposes to create a price based feedback control loop mechanism between an ISP and it's users. For the ISP, it computes the TDP (Time Dependent Pricing) prices to balance the cost of congestion that occurs during the peak periods while offering lesser prices during less congested periods. While this may solve the problem of congestion, it doesn't further the cause of pricing or valuation of user information.

With the rising amount of data analytics being performed on data on such a regular basis, [135] puts forth an idea to perform affordable analysis on expensive data. According to the paper, some analyses can be informed on inexpensive versions of low quality information because a small margin of error can be tolerated when large scale analysis is performed on such low quality information. According to this paper, this type of information can be purchased in versions which can be inexpensive for various data analysis purpose. Though it doesn't specifically use the term "data brokers", the entities from whom such type of information can be purchased can be assumed to any sort of a third party and not the user. Thus this justifies and reasons the analyzing of low quality information from the buyer's perspective and not the user's perspective.

To tackle the issue of picking the right candidates in a crowdsourced environment

and paying them, [82] develop incentive mechanisms for selecting the individuals to build a team and determine each individual's payoff for working in the team. They define a vector that specifies the skill level of the participating individual for every task. Their incentive mechanisms are - optimal mechanism (that chooses the cheapest team), greedy mechanism (that chooses the minimum cost per skill for a team but pays the asking amount for the individuals in the team), VCG-based mechanism (based on the Vickrey-Clarke-Groves method) and TruTeam mechanism (that picks individuals for a team based on greedy mechanism but pays the individuals the highest cost mentioned in the team).

[150] looks at measures to allow crowdsourcing of tasks in a truthful manner without sacrificing any utility. The current models for crowdsourcing work on the offline model where participating users report tasks that they can complete and submit their bids in advance and the crowdsourcer selects those users with low bids after collecting information from all the crowdsourcers in order to maximize utility. In this paper, the proposed mechanism allocates the tasks only to the user of the marginal density is not less than a specific threshold computed using the previous user's information and the budget for the current stage has not been exhausted.

[148] targets the smartphone user market with the aim of utilizing information from user's smartphones by designing incentives to attract users to participate in a crowdsourced endeavor. They provide three models - Threshold based auctions (that aims to maximize the utility of the model by using the first set of users as a threshold and do the actual selection from the next set of users), Truthful online incentive mechanism (that aims at truthfulness and builds on the threshold auction) and Truthful online incentive for arrival departure mechanism (that assumes that the users have specific arrival and departure times and must report the same truthfully).

[80] proposes a usage based dynamic pricing mechanism for the smart grid in a community environment that enables the prices to correspond exactly to the electricity usage in real time. [114] designs incentive mechanism for the users in a smart grid in order to get them to share their information. They utilize the Vickrey-Clarke-Groves

(VCG) mechanism to maximize the utility functions of a set of users. The users' reward for participation is a possible discount on the electricity payments. The emergence of the smart grid has led to research about how to utilize the information from the users to design efficient scheduling and pricing methods. [113] looks into the concept of pricing for the smart grid by enabling interactions between the smart meters at the user's side and the energy providers. The proposed algorithm based on the concept of utility maximization, attempts to find the best energy consumption levels for each user (also known as subscriber) to maximize the aggregate utility of each subscriber.

[49] provides a mechanism design approach to the issue of revenue management by elaborating on the existing models and explaining the possible extensions for revenue management for different scenarios. To address the issue of online pricing where we have users (called as agents) arriving over a period of time and the model is not appraised on their arrival before hand. [46] provide modified mechanism design approaches by elaborating over the issue of pricing wifi at starbucks. They use the existing mechanism design schemes like the VCG mechanism to address the problem and then discuss how the payments can be arrived at. Price discovery is how buyers and sellers come together and reach a suitable price for goods in the market. [60] analyzes the significance of price discovery and the effect of market competition in the process of price discovery. The presence of competition in the market enables the discovery of optimal prices. Mechanism design differentiates between various market mechanisms in order to arrive at the optimum mechanism for market economies.

[151] expounds on the conundrum of pricing a stock of perishable good by considering the dynamic pricing model based on the Markovian decision process for revenue maximization as a possible approach. Based on their analysis, they observe that at any given time, the optimal price keeps decreasing as the inventory decreases.

[68] develops pricing models and schemes for the Internet of Things that can be directly applicable, comprehensive and executable for the scenario of the Internet of things. The authors build and establish an ontological model using a combination of the W3C Web Ontology Language (OWL) and the Semantic Web Rule Language

(SWRL). The development of a similar type of model for pricing information is needed that can be utilized by the user himself rather than depending on a third party to do this for him.

For the traditional market with tangible goods and for regular markets dealing with regular goods, [142] mentions about the concept of Full Information Product Pricing (FIPP) networks which is basically complementing the price of the product being sold along with information about how the product was produced and distributed. This information is relayed across using trusted and certified parties. The authors feel that promoting such a framework will facilitate a better relationship between producers and consumers. We feel that this can and should be extended when dealing with information as an economic good to enable and foster the presence and growth of the information market among the producers and consumers of information.

This section describes the various techniques researched to look at pricing computing related services and products which look to increase an organization's revenue by reducing congestion in a grid, employing price discrimination, pricing cloud resources etc. While they may revolve around the concept of information gathering and analysis, they do not tackle the complicated task of monetarily compensating the information producers, i.e. the users.

2.5 Privacy concerns and pricing

Today with the increasing number of attacks and realizations about the breach of user's online privacy and personal private data, maintaining control over one's private and personal data should be a top priority for all internet users. With increasing usage of internet based computing and more recently cloud computing, we store all of our data, private or otherwise, online. Add the strong presence of social networks, we now have a person's entire history on the internet, up and running 24/7. There is a continuous ongoing tussle amongst different stakeholders interested in getting hold of the vast amount of user data. This is because access to this information gives amazing insight into a person's psyche and can help predict that person's behavior

and sometimes even manipulate that behavior.

There are numerous concerns about data invasive applications. This concern has led to the research community in developing different techniques to satisfy the different stakeholders involved though none seem to have been commercialized as yet [43]. [90] has targeted the invasion of mobile apps on teenager's phones and the lives. The privacy concern about this encroachment has found echoes amongst many [30]. This has prompted discussion and research work to allay any concerns, users may have about their private information [81].

Now, for a user in a social network, the privacy concerns for him or her with regards to private data (either about the user or related to the user) can be, the *lifetime of the information* (how long can someone access this information), the *transitivity of access* (who all can access this information) and *claims-based access* (for what purpose can a person access this information) [110].

With the increasing usage of the internet by most users for not just social but also for commercial reasons, consumers are getting increasingly wary of the potential misuse of their information and of losing vital control over their information [20].

Understanding user's privacy and information sharing behaviour is complicated and the economics of privacy even more so. Though users have stated numerous concerns about the potential misuse and abuse of their private and personal data, they still continue sharing the same on social networks and for other e-commerce transactions. This is said to be the privacy paradox where despite being aware of the implications of their actions, users do not choose to exercise privacy preserving options for their information. The privacy settings on online social media can often be difficult to navigate. Researchers have tried to analyze the privacy settings of Facebook in an attempt to investigate how effective it is in protecting the privacy of the users [83]. Their focus in this paper has been on highlighting the glaring discrepancy between the expectations of the users and the actual technical reality of the settings provided. Through the means of the survey, the authors conclude that despite the privacy settings in place quite a lot of the content is shared anyway. There

is a large gap between what the users expect and what is delivered to them from the point of view of privacy settings. Though this was conducted almost three years back and a lot has changed, this just adds to the belief that social media sites do not do much to protect the privacy of the users as per what the user thinks or believes.

In an empirical study conducted by [23], it was found that participants choose to opt for lower price despite having to divulge more personal information. Also, keeping all things equal (price, product, service etc), more privacy friendly measures did not find more favor amongst the participants. It is this apparent apathy and unwillingness to pay that has baffled privacy experts for ages. [10] looks into this difference in the user attitudes by applying concepts from behavioral economics. They show that expecting the individual user to behave like a rational user is highly unrealistic by applying models of 'self control bias' and 'immediate gratification'. They conclude that users cannot always be trusted to make the best privacy decisions about themselves. They state that a combination of privacy technologies along with increasing awareness among users could help in the privacy welfare of the users.

Another arena is the investigation and research of the obvious difference between privacy attitudes of users and their actual behavior [15]. This difference being the apparent willingness to give up on their privacy for adequate compensation and the resistance to the usage of privacy protecting measures. The authors conducted a survey and came to the conclusion that the privacy preserving model developed cannot be based on strict rationality since users at times may not always take rational decisions due to various factors.

We agree with the authors on this and feel that more flexibility must be given to the users to decide for themselves how much privacy do they wish. But we strongly feel that users must be given complete information about how their data is being used so that this can enable the users to make a more rational decision about the release and subsequent usage of their information by third parties. One of the possible ways to build and improve consumer's trust and alleviate their privacy concerns is to have a privacy policy in place. Most online businesses have these policies on their website

and are usually built around the United States Federal Trade Commissions five widely accepted principles of fair information practices: notice, choice, access, security, and enforcement. A study conducted [144], aimed to understand the relationship between the privacy policies of businesses and its effect on consumer trust and privacy concerns which reflect the consumer's willingness to provide personal information. The results reinforced the importance of having an intelligible and easily accessible privacy policy for users to understand across borders.

There is always a conflict between the idea of personalization and the need for collection of user information to cater to that personalization. This is elaborated upon in [26]. Information about the users (or customers) allows the organizations to target them with specific advertisements catered to their likes and dislikes. It also allows them to target the users with tailored products with specific prices. The implementation of privacy is seen as preventing the above and as such is seen as being detrimental to the process of personalization. This paper shows that in the presence of perfect competition in the market, this negative connotation is unfounded because the organizations and vendors offering privacy enhancing technologies end up making a higher revenue. They also state put forward the idea to have a more flexible idea of privacy measure rather than just a binary measure.

Most of the data available over the internet today, is data collected about users and their online behaviours. This data is more often than not by organizations to provide'personalized' services or advertisements to users in order to better benefit them. But though these personalized touches are sometimes useful, they also give reason for privacy concerns. A recent study, [133] conducted on the above subject of personalized services, broadly categorizes the concerns about them into three domains - social networks, behavioral profiling and location based.

In the position paper by Gummadi et al [55], the authors expound in the problem of privacy in online social media these days. With the amount and nature of the information being shared through these websites and the number of people who can gain access to this information, it is up to the user entirely to control and protect

his or her privacy using the various controls provided on these websites. But as the authors argue, these controls are in themselves so complicated and confusing that users do not end up using any of it at all. There is no single simplistic solution to protect your privacy. The important thing that the authors have failed to mention is the protection of the private information from the websites themselves.

[93] mentions the importance of user information not just in the social media but especially in the grid, which is soon to be common place. This paper describes the use of additively homomorphic encryption in order to preserve the privacy of the users in the grid. While protecting the privacy of the user, this work aims to provide up-to-date, accurate and aggregated information about the electricity consumption of the users in the grid to the electricity supplier for analysis.

Private information retrieval solutions allow the retrieval of addresses of information only. But for users using the cloud computing environment, this is not available without huge computation and communication costs. The method described in the paper [147] achieves the trade off between cost of retrieval and the degree of privacy. It allows the user to design the level of privacy that they are comfortable with and with the user budget.

In [134], the amount of information gleaned from a statistic about a user is quantified. The goal of this paper is to provide a method that can be automated to determine the amount of information the flows through any program that performs the task of computing statistics. This paper assumes that the adversary in this situation would like access to the coveted distribution statistic that is derived from some function over the input distribution.

[139] tries to detect the presence of economic motives behind the individual concerns over privacy. It suggests that all type of information must be treated as equal and that any type of control over information should be exercised on the usage of the information.

There are different policies that could be set up as a guideline of sorts for private data in social networks. One way to go about this is the joint enforcement of privacy

policies for shared data on social networking sites [125].

Privacy is a complicated issue. Every user has a different take on the amount of privacy acceptable to him or her. Privacy economics attempts to use various methods like market mechanisms (from economics), technology (computer science) and policy (law) to come up with a satisfactory balance amongst stakeholders (mainly individual users and organizations). It tries to quantify the perceived benefits when user's private data is shared and when it isn't. The latest in this fledging field of research is something called 'Soft paternalism' [11]. The idea here is to create and construct systems that gives users the option of making informed choices about their private data; an attempt at giving user's more control over their data.

The idea of using differential privacy to enable buyers to buy and have access to information [50] puts the buyers at a backfoot since he or she will have to shell out a lot of money for any usable information. Differential privacy [42] helps to define a meaningful way to bound how much a person or individual could potentially be harmed because of the loss of his or her privacy. Thus, the output of the query is a randomized answer such that even with some small change in the data base does not change the distribution of the output significantly. The catalyst behind the origin of differential privacy was the need to be able to estimate and calculate and subsequently release some information of a sensitive or private nature from the available database. One way of defining privacy is by stating that after obtaining the answer to some private query about some individual, we do not gain any extra knowledge about that particular individual. The challenge is then, to retain some utility over the result, it is inevitable that we must allow for some information to leak out. What we can control, if the amount of leakage that occurs i.e., we can set a bound over the leakage.

Game theory [71] has also been proposed to allow users to avail of a pay off. But they only mention certain specific scenarios and only if the users are the 'winners' in that game. In their study they make use of the concepts of differential privacy and mechanism design to develop auctions for keeping in mind the privacy constraints and compensating users for their privacy loss. In this the users participating in the auction

state a privacy budget or a constraint that depends on their utility. Depending on this, the auction mechanism chooses the ideal user and compensates them accordingly.

The problem with users mentioning their privacy budget out in public is that, it leaks out the important detail of how valuable or private a particular piece of information is to the user. Hence a recent work, [105], focuses on remedying this by allowing users (or agents) to specify an upper bound based again on how it would affect their utility rather than a privacy parameter. To encounter high privacy valuation upper bounds, the authors have established a threshold, letting those with lesser valuations benefit and still maintain the privacy of all participants.

In today's age of the pervasiveness of many devices in user's lives, privacy and private data management is a major concern. The fact is that as more and more data is shared online by users despite knowing the risks, there is the real possibility that these users may not be appeased by the presence of privacy control measures without understanding the cost of what is at stake here - their information, which has immense value as a commodity in the internet information market. It is here that understanding the valuation of their information will help users to realize the potential and the importance of their information. This could enable users to exercise better control over their digital footprint and leverage it for monetary benefits.

2.6 How valuable is user information

Advertisers monetize on user data on social networking sites, based not only on their profiles but also on the user posts. The main issue with this is to identify 'monetizable' posts [102]. For this, patterns are identified based on word choices used in the posts and used to display specific targeted personalized ads for the users. The main point here is that the algorithm is based primarily on user posts related to queries about a product or a service. The data used in the algorithm were not private and nor were the intents of the users to keep it private.

A study carried out by Aricent and Frog Design [18], (see Figure 2.2) revealed the worth of a user's information in the information market.

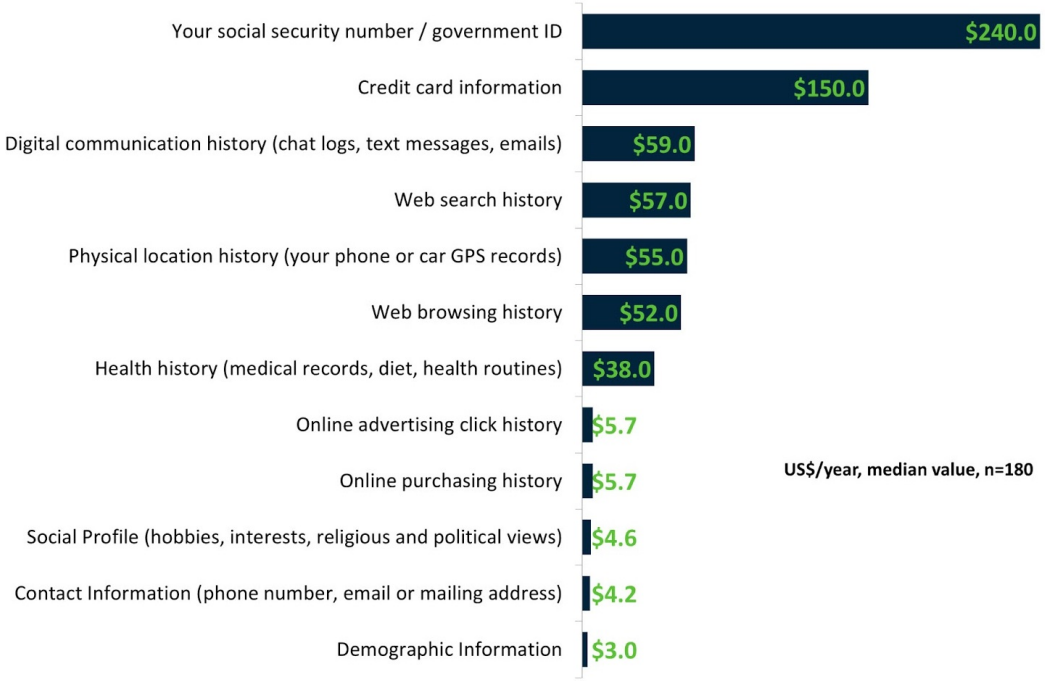


Figure 2.2: How much is the user's information worth?
(Source: Aricent/Frog Design, primary research 2011)

Another avenue for advertisers and organizations to profit from user information, is to enforce something called as "price discrimination". By this, depending on user valuations gained from user information and user data collected (with or without their knowledge), items are priced differently or at times items matching a user's profile and price valuations are displayed to the user. The evidence of price discrimination is present all around us as shown by Mikians, Gyarmati et al [98] [99]. Based on their findings, they have proposed the creation of a system that allows users to realize when such price discriminations occur.

Auctions are also a way to allow users to sell information. [111] uses the exponential mechanism to auction off user information. The users decide what personal information about themselves they would like to sell in the market. But in an auction, the price depends on the number of interested parties and how much they are willing to pay, which puts the users at a disadvantage. But we feel that though the exponential mechanism is used, there is no way we can ensure that the data values are not being manipulated by the information aggregators (like data brokers) for their own benefits.

Increasingly we hear of the issue of organizations and firms capable of tracking and storing consumer information selling it like a commodity in the so called information market to third parties. One way of looking at this is that depending on the total profit obtained from a particular good or service, a firm can decide whether to collect or not collect information about the consumer [62]. The information collector in their scenario is an internet service provider (ISP) which has the ability to keep a log of consumer's information along with the websites he or she has visited making a complete profile of the consumer and perhaps sell this information to third parties who would then use this information to provided targeted advertising and personalized services tailored depending on the type of consumer. But this comes at a loss of privacy to the consumer and with no additional benefits except being bombarded by more targeted advertisements. Based on this concern, the authors suggest that firms collecting information provide consumers with the option of either letting their information be collected, in which case they will be suitable compensated in the form of some incentives or not collect the information at all in which case they cannot benefit off any incentive.

It is not always easy to provide users with the level of privacy and anonymity they seek [12] . The authors analyze the cost associated with providing anonymity by the creation of nodes which forward the messages passed on to them as it is i.e., without breaking the anonymity. The costs associated are the costs associated with any computer network (maintenance costs, computing costs etc). Each agent in this system has a certain reason for wanting to remain anonymous and has a certain value attached to it. The problem here is that anonymity depends a lot on the honesty of the node, the amount of traffic generated and the associated costs.

Coalition games (part of game theory) can be used to come up with a method to compensate users for the loss of their privacy [71]. The paper discusses and elaborates on two scenarios, one involving a market survey and another involving something akin to a recommendation system. In the market survey, individual users state their preferences with the product or service having the majority being available. Here the combinations supporting the majority increase the pay off of the vendor and

subsequently individuals also profit if they are in the majority. The authors use the concept of the core and the Shapely value to prove the same. On the other hand, in the other scenario, based on the Shapely value, the more novel and unique the contributions, they end up receiving a better pay off. But this study targets specific scenarios and only benefits the users if their answers in the scenarios adhere to a certain criteria.

As far as pricing data goes, though the norm seems to be to buy from a set of predetermined database with no flexibility with regards what the buyer actually needs, researchers [72] have devised a new scheme to price data in their setting in a relational database. According to their scheme, called as "query based pricing" , based on the fundamental economic pricing concepts of arbitrage freeness (where there is no taking advantage of varying prices of different but possibly related items or in this case queries) and discount freeness and the price fixed for a certain dataset, the price of the query for the dataset the buyer specifies, can be derived easily. This work is however, bereft of any mention of the handling of the privacy of the person to whom the data belongs. And as mentioned by the authors, this work seems to fundamentally be linked with the data being a part of a traditional relational database with a predetermined price fixed for each query. No mention is made of how this price is arrived at. And most importantly, this does not benefit the users whose information is actually being stored in those databases.

Another technique to sell private data is by linking the price of the data with the amount of noise added to the data [79]. To achieve this, there are three components in the pricing model - the users (who have the data), the buyers (who are interested in the user's data) and the market maker (a third party who is a link between the users and the buyers). The model is simplistic enough - the market maker collects the user data and depending on the amount paid by the different buyers, adds the appropriate amount of noise and sells them the data. The market maker is also responsible to make sure that each user is compensated accordingly. The problem we visualize here is with the valuation of data. The authors do not really have a set mechanism design in place to appropriately value the data. Also the market maker, can be prone to

act maliciously especially since he or she has the unperturbed data with him or her. This similar technique has been mentioned by [52] where a third party (or a middle man) as a facilitator is mentioned to decide pricing. We feel this is risky because this facilitator is entrusted with the user information, who could use it in a malicious manner.

An attempt [19] has also been made to quantify location information obtained from mobile networks. This information is pretty valuable for organizations catering to location based services. The authors have developed models that include parameters like geographic location of the users, their requirements and their ratings of the services used. Though not exact, these give an estimated revenue that can be generated using this information for these services.

Apart from the vast amount of personal information available online, another valuable piece of information for organizations is the browsing behaviour of the online users. This is what is mentioned in [34] as they go on to explain the importance of this browsing behaviour for big conglomerates like Google and Facebook that provide tailored ads and services based on a user profile that is created using this type of information. The authors have attempted to try and understand how users value their own information using a survey based technique. Their results state that though users do not like the monetization of their personal information, they do not mind collection of their browsing behaviour. These users do not seem to mind giving up their information for usage of services but dislike if their information is collected or used for additional revenue generating schemes.

A study conducted by *The Organisation for Economic Co-operation and Development (OECD)* [106] looks at the idea of measuring the monetary value of personal data. According to the study, the way that this personal data is used in different situations has a big role to play in determing its value and thus have presented a comparitive analysis of the different methodologies that assigns monetary value to personal data. One of their approaches looks to determine the monetary value by delving in to the market capitalizations or revenues for every single record for those

firms whose entire working and business revolves around personal data. The other looks at examining the cost of a data breach. And another approach is to conduct surveys to understand how much individuals value their personal data.

2.7 Summary

The concepts of price discrimination, price bundling, dynamic pricing are all aided by the information gathered about the users. These practices are unfairly biased against the users and leverages the user's information against them. For example, the information category of location might classify a user as "wealthy" if he or she comes from a upscale neighborhood and thus price a product higher for him or her or possibly combine it with another product. While the recognition of information as an invaluable resource is seen all the time, especially with businesses going to great lengths to acquire information about not just existing customers but also other masses, steps taken to compensate the actual producers of this information, i.e. the internet users have been relatively rare. As seen in this chapter, the idea of pricing information is still nascent and we believe the pricing models proposed in this thesis will be a foundation in this field.

3

Overview

Big Data has become an ubiquitous term used today, across all organizations and businesses. The typical characteristics of big data: volume, variety and velocity, render it almost beyond the capabilities of a traditional database and have brought about a revolution in the way data mining and data analytics have been performed [41]. But, of all the importance attached to information, the assessment of its value has gone largely unnoticed.

No information about a user is useless. Millions of little pieces of seemingly useless and random pieces of user information are put together to build a 'profile' of sorts about the user that shows his behavior, his thinking and can possibly predict his future actions.

The figure 3.1 shows us how seemingly ordinary pieces of user information can

Description	Data	Big data
Descriptive	Age, gender, income, demographs	Attitudes, psychographics
Social	User-defined	Influence peers
Location	Addresses	Real-time locations
Interaction	Next-available agent	Based on the personality of the customer
Relationship history	Transactions, Employee interactions	Operations usage, Data-driven interactions
Next Action	Resolve issue	Among the 1,000 potential offers, make this one

Figure 3.1: Insights from user information

reveal amazingly deep insights about a user and his behavior which can reveal the thinking and predict the future actions of a user.

The value of user information for a decision making entity in a variable setting can be said to be linked to the behavior and relationship with similar decision making entities. In our case the decision making entities are the organizations and conglomerates who leverage the user information and the corresponding insights for better decisions that lead to better profits. The market for personal data is steadily growing in recent years. This has led to increasing amounts of time and money spent by conglomerates towards gathering maximum information about the users. This leads us to believe that users today have the chance of making money from their information that they generate and create everyday.

In this chapter we present a general overview of the work presented in this thesis. We explain the catalyst for this research work by the form of a brief case study that we conducted with the normal ordinary internet users.

3.1 A case study about user outlook regarding information

To better understand how ordinary everyday users think about their information and to see if they would be amenable to the idea of monetizing on their digital footprint by selling their information, we surveyed around 50 users from diverse backgrounds and age groups and asked them various questions regarding what they think about the value of their information.

Figures 3.2, 3.3 and 3.4 show us the concern that users have regarding the sensitivity of sharing various types of information. We asked the users to rate their concern from a scale of [1-5] with 1 being the least concern and 5 being the maximum concern.

From the figures we can see that the most concern is for information categories like sexual orientation, medical history, credit and debit card feeds, salary, location and contact list. These are categories that are considered as private and personal. The concern for the categories of social media feeds and medical history is more spread out with some being more concerned than most.

Apart from the concern for various information categories, we also asked the users to select the organizations to whom they were most comfortable with sharing their information. Figure 3.5 shows us that most users are comfortable with sharing their information with educational institutions. The reason for this could be that educational institutions are not seen as having any sort of malicious intent. This is followed by governmental organizations and media organizations (like Facebook or Google). The reason behind this willingness could be that governmental organizations are seen as authority figures whose need for user information is indirectly or directly tied to the user's benefit (like security or providing better services) and media organizations like Facebook or Google provide those services that the user has become completely dependent on and familiar with and hence may feel beholden to share their information with them. On the lower end of the spectrum are banking organizations and advertising

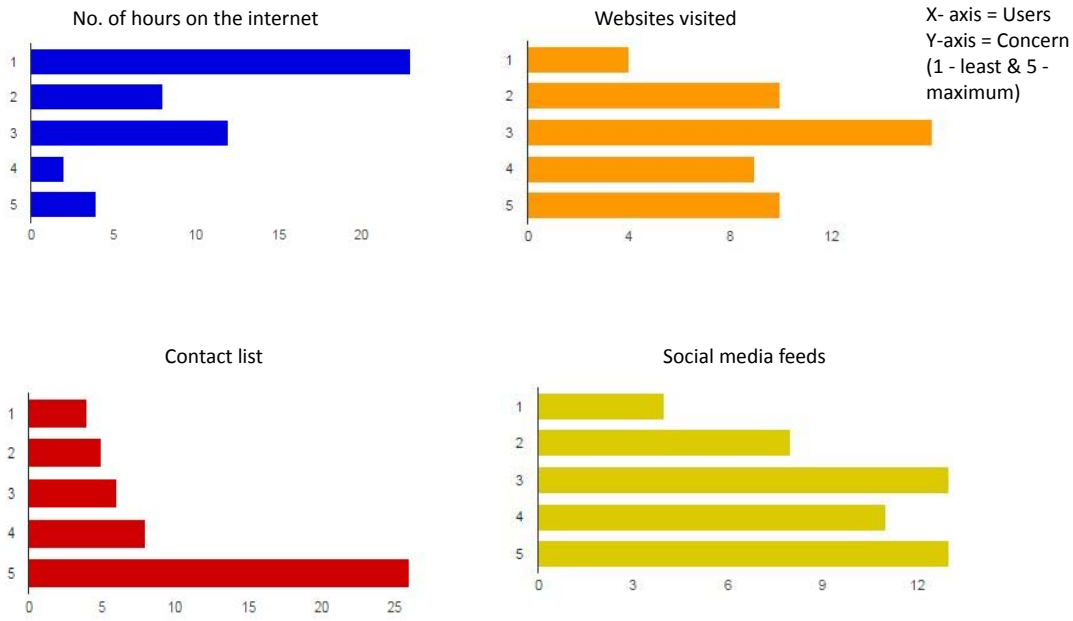


Figure 3.2: User's concern about sharing their information (a)

Figure 3.3: User's concern about sharing their information (b)

Figure 3.4: User's concern about sharing their information (c)

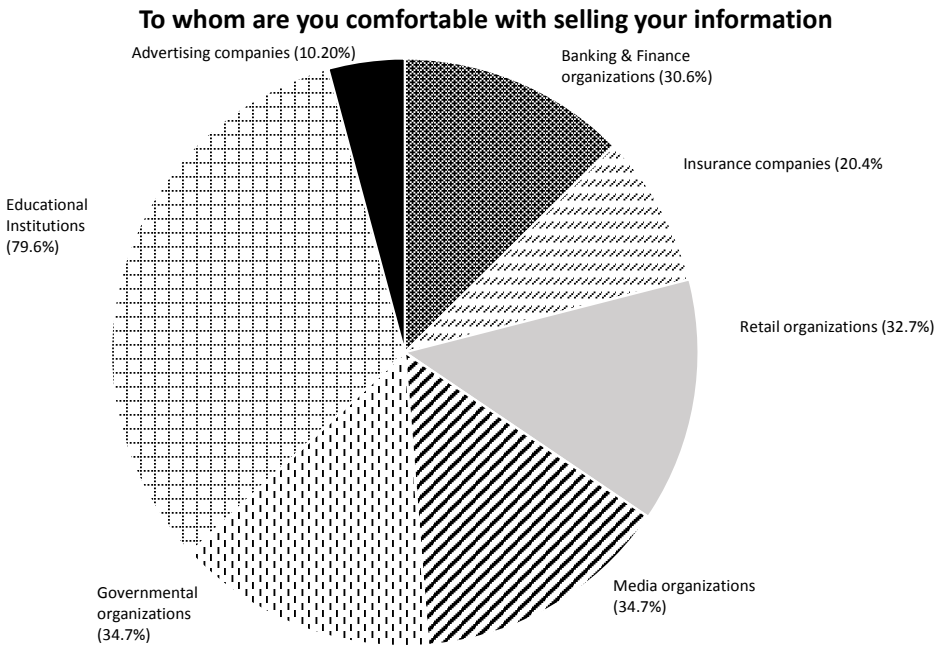


Figure 3.5: Whom are user's most comfortable sharing their information with

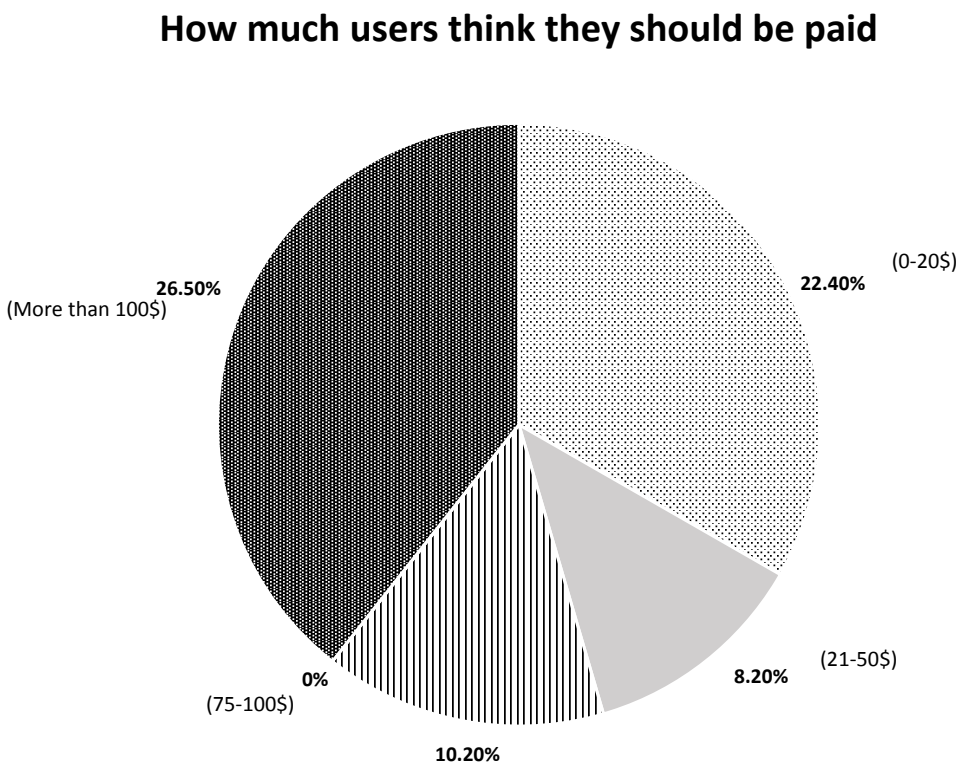


Figure 3.6: User's opinion about how much they should be paid

companies with whom users are more reluctant to share their information. This is probably because these types of organizations are seen as pure profit making entities who exploit any available snippet of user information for their own benefit.

We also inquired among the users how much did they think they should be paid for their information. From the figure 3.6 we can see that maximum users tend to think of payment in terms of extreme numbers i.e. they either choose the minimum or maximum. This is because users tend to undervalue or overvalue their information because they lack a firm basis of prior pricing precedent to guide them. The question then is to get them to accept and visualize a middle ground which is reasonable and fair to the potential buyers of this information.

But as figure 3.7 shows, most users are acceptable to the idea of compensation in the form of other formats like discount vouchers, store credit etc. This shows us that the idea of monetary compensation need not be tied to actual money but could be incorporated by the potential buyer into some part of the potential buyer's business.

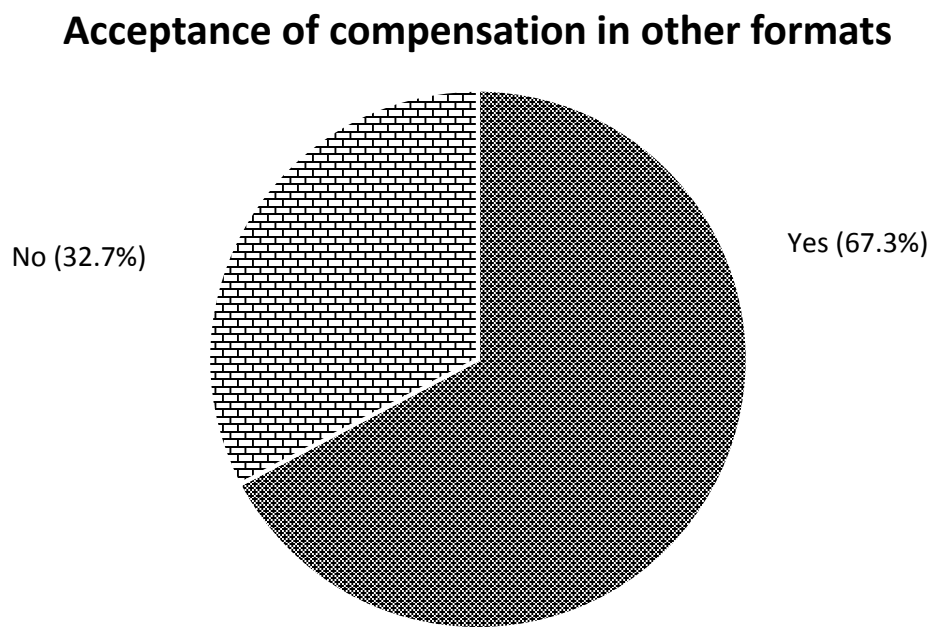


Figure 3.7: User's willingness to accept compensation in other format like discounts, store credit etc.

The analysis of the above results showed us that:

- Users are able to objectively consider their information as an asset or good to be sold in the scenario of an internet information market
- Instead of direct cash for their information, users are willing to accept the compensation for their information in other formats like discounts
- Users have some hesitancy or concern regarding sharing some information categories or types, mostly of a personal and private nature
- Despite the above concern, with the right incentives in the form of monetary compensation, the users could be convinced to sell their information. These right incentives could be a *guarantee of privacy*, an assurance of *monetary compensation* etc

The above case study helped us understand that users would like more control over how they can share their information and with whom. This leads us to believe that they would prefer dealing directly with the potential buyers rather than any middle

party or conduit when it comes to leveraging their information as an asset in the internet information market.

3.2 Research description

Based on the analysis of the above case study, we have proposed different scenarios that depict the pricing models for user information corresponding to varying circumstances. Since the users have some hesitancy over sharing or selling certain types of information, we have taken the following method of research for our pricing models:

- With obfuscation to assuage user's of the privacy of their information for those categories of information which they may have some skepticism over sharing. The obfuscation models allow for masking of user data without too much distortion of the actual information and the compensation is adjusted according to the obfuscation levels. This is presented in Chapters 5 and 6 with different measure for estimating the quality of information.
- Without obfuscation for those information categories that the user has no hesitancy over sharing. These are presented using different theories and different valuation concepts in Chapters 4, 7 and 8.

The rest of this section is devoted to explaining the stakeholders in our models and the research assumptions that are the building blocks of the model.

3.2.1 Stakeholders

The stakeholders in our scenario are chiefly the buyers and sellers who are involved in the transaction of the sale of information.

The buyers are the organizations and businesses interested in information either for direct purposes (like to target new customers or market research companies) or indirect purposes (like retail companies to track sales). They need this information to analyze and understand to get into the minds of the people in order to pump up their own profits and sales. These buyers are interested in all sorts of information about the users that can help them detect patterns in user behaviour that they can leverage for their benefit. Currently, these buyers pay money to data brokers to purchase this

information about the users for which ofter they have to end up paying a lot of money.

The sellers in this transaction are the ordinary internet users who generate information about themselves and their behavioral patterns while they use the internet and their smart phones in their daily lives. They can be said to be "potential sellers" since these people do not know the potential and value of this information that they are generating. And since they do not realize the value and importance of this information, they do not particularly care who has access to this information and for what purpose. The one thing to jolt this apathy is to inform these users about the monetary loss they suffer when this information is taken without their knowledge. The objective of these users should be to be able to obtain monetary compensation in exchange for access to their information.

3.2.2 Research Assumptions

In the preceding chapters, the pricing models discussed have handled different data types - both numeric and non-numeric.

For the purposes of this research, we have taken into consideration un-encrypted type of values. The reason the user information needs to be unencrypted is because encryption introduces changes to the information values and thus reduces their utility in terms of usage for business and analysis purposes. Our research is looking into the idea of information pricing only and as such we have not incorporated the idea of encrypted information values for the purposes of our research.

The type of information which we have discussed belongs to the user. This could be the information about them (like salary, home co-ordinates, age) and information that is generated by the user (like credit card expenses, monthly savings, GPS co-ordinates, number of hours spent online etc). For the purposes of this research we have assumed the integrity of the information is satisfied, i.e. the information that is shared by the user is about the user and is the actual information that he is sharing.

3.3 Summary

In this chapter, we presented our case study and the result analysis from that case study to gauge the user's idea about selling information like a commodity in the internet information market. From these results, we were able to understand the user's concerns and have worked on our pricing models accordingly. This chapter also describes the research models which are explained further in the forthcoming chapters and our motivations driving the different scenarios for which these models were developed. The scenario for pricing information is quite realistic and with increasing awareness will be the necessity of the future. Having appropriate pricing mechanisms for the same is the way ahead.

4

Pricing based on Information Loss Measure

In this chapter, we present the pricing models that correspond to the scenario for users to estimate the value of the information that the users will choose to share with the prospective buyers. This information value is then used as the basis for the calculation of the price for the information of the users.

When the information about a user is known, it can be considered as a loss for the user and a subsequent gain for whoever has possession of this information. This is because the information reveals aspects of the user that the user may not want to make public. This information would also convey a lot about the user to the other parties. The pricing model presented in this chapter looks at compensating the user for this loss by first calculating the information value using the idea from Shannon's theory and then keeping in mind the idea of a fair information market, this information value is combined with the demand in the market and thus the price for

the information of a user is arrived at.

In this chapter, we have presented two pricing models - with regular demand gauging and with exponential demand gauging. For the purposes of the pricing models, we have not implemented privacy preserving measures for the user information.

4.1 Introduction

The concept of Shannon entropy [117] has been used widely in communication system studies to understand the privacy and anonymity problems that arise. It is used mainly to calculate the amount of uncertainty for an outside bystander when it comes to figuring out role assignment (sender or receiver) in communication systems. Shannon entropy is used to assess the extent to which an outside bystander or observer would need additional information to classify the roles assigned in a communication system. In [24], the authors use Shannon information theory along with Option pricing theory to value privacy. But they look at it from the point of view of option pricing and use it to determine the amount of privacy lost for a user.

We believe that like any other good in the market; in the information market also every information type or category, has different demands in the information market. Eg: The information category of 'salary' could be more highly in demand than 'educational qualifications'. To understand this demand, we would require input from the interested buyers to signal the most sought after information category in the information market. So following the simple economic concept, a more in demand item, in our case information category, would be priced higher than the one in lower demand.

But just the demand is insufficient to determine the price. When a user reveals his or her information (or sells his or her information), the user needs to understand the value of the information that has been lost. We aim to calculate this value using Shannon's information theory.

4.2 Pricing model with regular demand gauging

In our models, we utilize the concept from Shannon's idea of amount of information by understanding it as the amount of insight gained when an attribute-value from an information category is revealed. As explained in [86], Shannon's idea is to understand the number of questions that needed to be asked to correctly guess the attribute-value.

The main idea driving this is the concept of entropy, which is the amount of unpredictability associated with a random variable or a process. So, according to the theory, the Shannon entropy H, which is the uncertainty associated with predicting the value of the message in a communication medium is equal to $H = -\sum p_i log_2(p_i)$ where p_i is the probability associated with an instance if the *i-th* possible value of the source of the communication medium.

While we have Shannon's information theory to guide us to understand the amount of information that has been divulged, we need to find a way to understand the demand in the information market from the buyers interested in buying the information from the users.

To understand the demand from the buyers, we ask the buyers to state the amount they would be willing to pay for a particular information category. Then taking an average of this 'willingness to pay', we determine the demand in the information market for a particular type of information. This is the regular conventional approach with the assumption that the buyers are *honest* and *non-colluding* who state their demand truthfully.

In this process, every information category will have its own demand price or *base price*. This is the guiding price determined by the information market. It informs us what the buyers are looking for, what type of information categories are currently in demand and convey to the seller a sense of price discovery. The higher this base price, it means the demand for that information type of category is high, and this means that, that information category or type should nett a higher price. This is the guiding price determined by the information market. It informs us what the buyers are looking for, what type of information categories are currently in demand and

convey to the seller a sense of price discovery.

Since the user now knows the nature of the information market and what type of information is sought, each user is free to choose which information category attribute-value they wish to divulge. They have more control over their information and the access to their information. Based on the discussion above and applying it for our information scenario, each user can get an idea about the amount of information divulged by each information attribute-value they share using Shannon's information theory as $H = -p_i log_2(p_i)$. But without knowing the information values from other users, it is not possible to calculate the answer to p_i, which is the probability of a particular attribute-value occurring in a population of participating users sharing their information.

For a continuous random variable (say X) with density *f(x)*, the continuous entropy is defined as,

$$h(X) = \int_S f(x) \log \frac{1}{f(x)} dx$$

for example, if we have a continuous range say $[a, b]$, then we can simplify it to say the amount of information gained is proportional to $log_2(b - a)$. Thus the user can easily calculate his information value for every information type.

For this purpose, we ask users to state the range in which their information fits in most closely. From this range obtained from a set of users, we determine the probability distribution for every range for every information type [94].

Coupling the idea of *base price* (which determines demand in the market) and *information value*, we can arrive at:
*information price = base price * information value.*
We calculate this for every user for each information category that the user is willing to share.

In this manner, each user can calculate the price for each of his information attribute-value and obtain a revenue equal to the sum of the price for each information attribute-value. Hence, $\boldsymbol{Revene,\ R,}$ can be obtained by $\sum \boldsymbol{price_{ivalue}}$ where $\boldsymbol{ivalue}$ is each *value of each information type that is shared.* Figure 4.1 shows an overview of

the process.

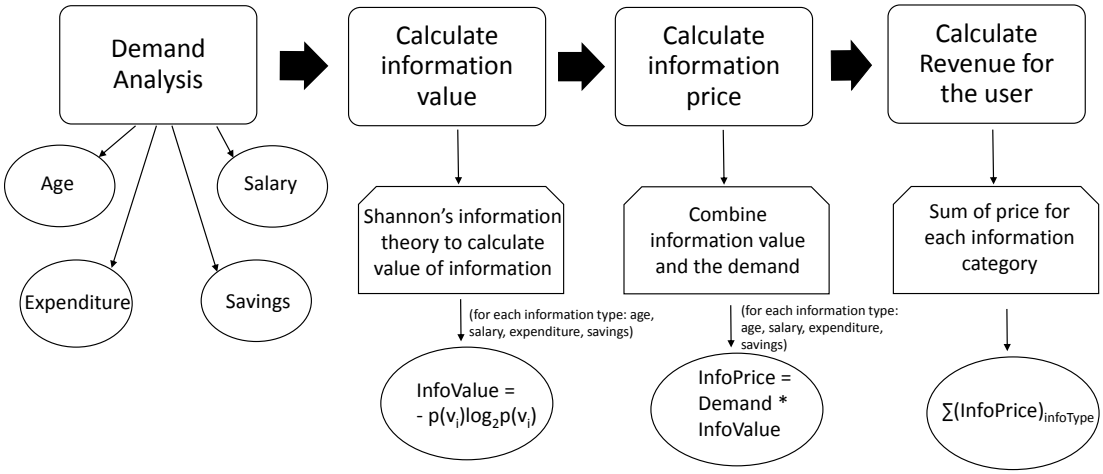


Figure 4.1: Overview of the pricing model

Pricing model with traditional demand

for each user **do**
  for each information category k **do**
    Calculate the information value for range $[\boldsymbol{a}, \boldsymbol{b}]$
    $\boldsymbol{value_k = -log_2(b - a)}$
    Obtain demand from buyers *demand*
    Calculate $\boldsymbol{price_k = demand_k * value_k}$
  Calculate revenue for each user $\boldsymbol{revenue = \sum price_k}$

4.2.1 Experiments

To obtain a better understanding about our model we solicited the help of 4 participants who took on the role of buyers and explained the underlying scenario and motivation for the pricing model and we asked them to state the amount they would pay per user for different information categories, as buyers who could make use of these information category attribute-values. Since with our model, currently we can work with numeric attribute-values only, we had to be specific about the types of information categories to be solicited.The information categories that they were asked were as follows:

- Age (scores or marks)
- Monthly Salary

- Monthly expenditure
- Monthly savings

We then get the demand in the information market for the various information categories. We took an average of the demand for each information type.

This was then presented to another set of participants, who took on the role of our users. We then asked these 20 participants, to state which information categories they would be willing to participate in, i.e. divulge their individual information attribute-values. Then we presented them with a range for each of them and asked them to enter the range that they felt their information attribute-value reflected best in. Using this, we calculated the *information value*, as described in the previous section, for each user for each information attribute-value.

From this, we calculate the probability of the user's value in the participating population of users. We then combine the demand and the information value to calculate the price for the user's information.

Armed with these, we were able to calculate the price for each attribute-value and the cumulative revenue each user stood to gain.

This demand was then shown to the 20 users who were asked to choose which information category they would like to sell or reveal their information attribute-value for. According to their wishes and comfort level, the users chose the information categories to reveal their information.

Our method of obtaining user data to test the efficacy of our model while rudimentary, gives us an effective baseline and delve further into the ideas surrounding information pricing. Our model can be scaled up to incorporate the "real world" data, that can be possibly obtained at real time from multiple sources, to enable pricing of information.

With the results from our model, we can acquire better means of obtaining data from a more 'real world' point of view.

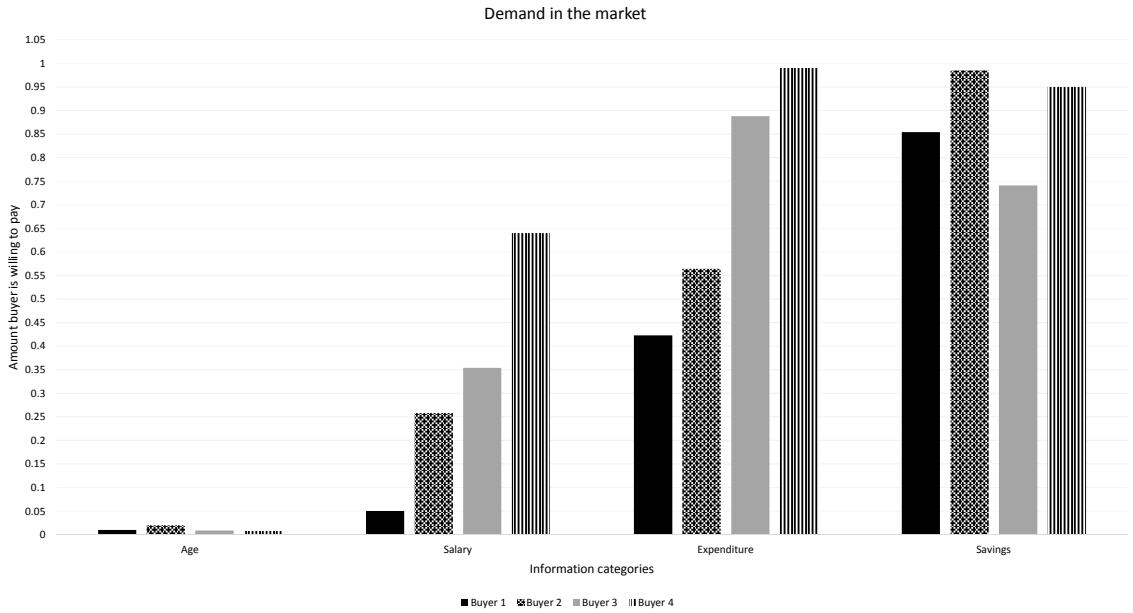


Figure 4.2: Average demand for various information categories

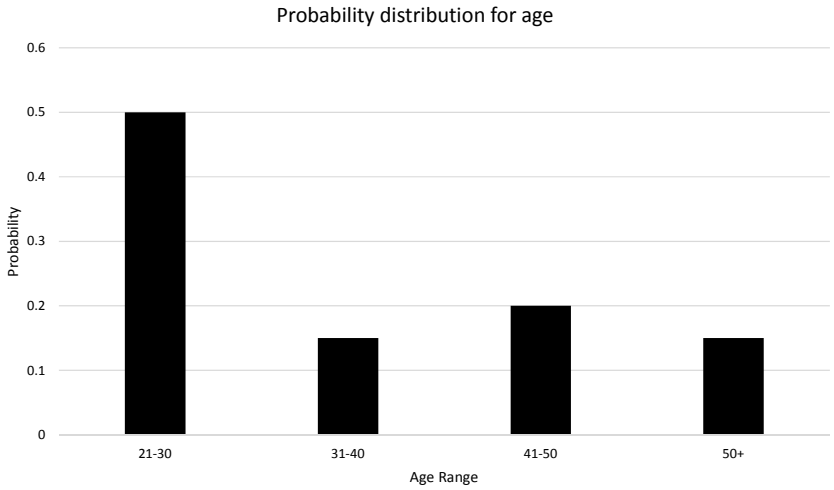


Figure 4.3: Probability distributions of age

4.2.2 Results & Discussion

Figure 4.2 shows us the buyer's responses for each information type. Generally for regular goods the demand can be charted based on the prices. But with a non - regular good like information with no precedence for the pricing, we have analyzed the demand by asking the prospective buyers to state their willingness to pay for the information category of their choice. After we obtain the values from the buyers ,

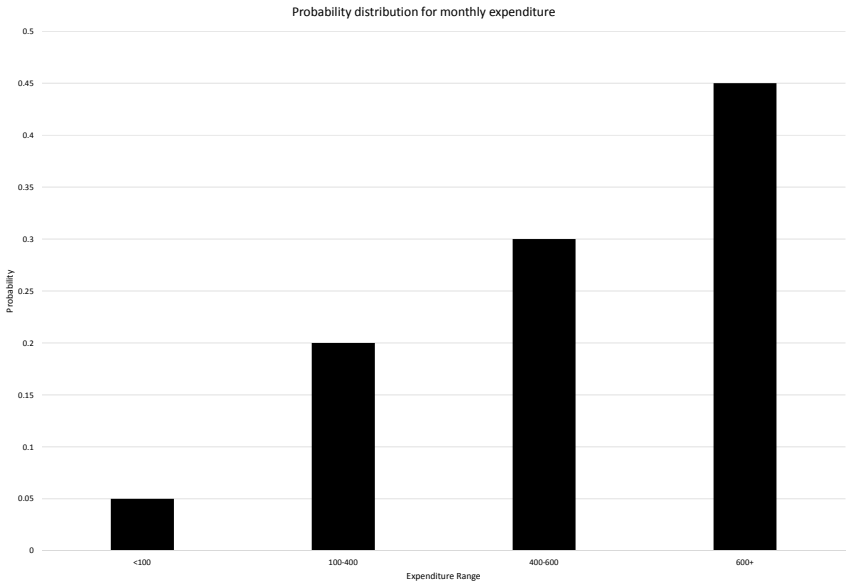


Figure 4.4: Probability distributions of expenditure

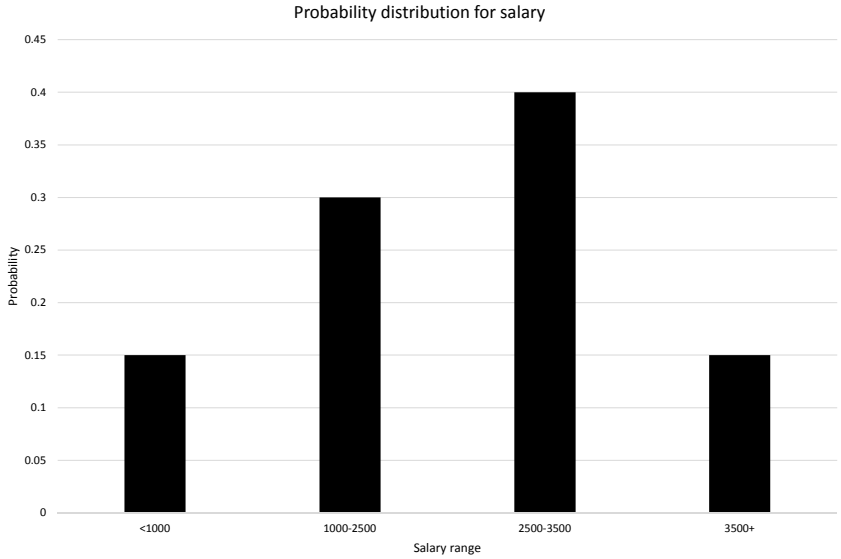


Figure 4.5: Probability distributions of salary

we average it for each of the information categories to obtain the demand for each information category. From the figure we can see that the demand for the information categories of salary, expenditure and savings are quite high.Savings especially seems to be in exceptional demand. This demand graph gives us a clear picture regarding which types of information categories are popular among the buyers in the information market. This gives us a solid economic foundation to base our prices on.

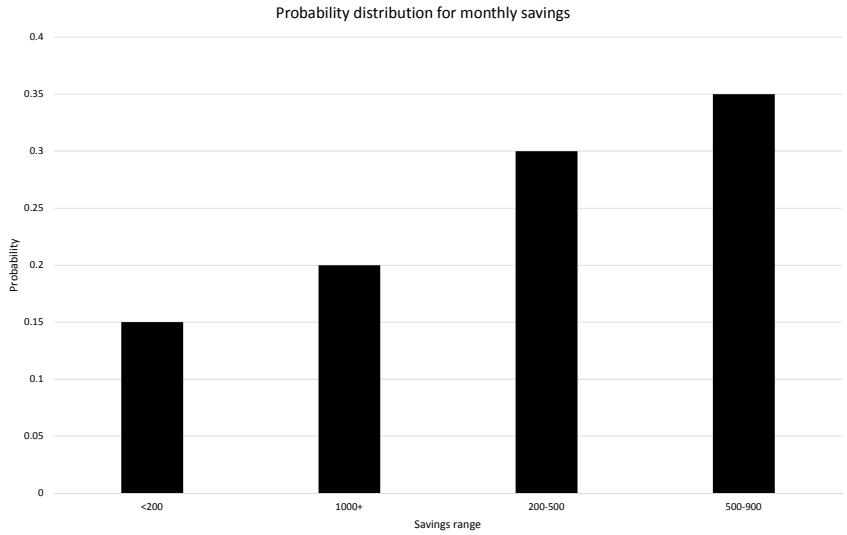


Figure 4.6: Probability distributions of savings

To establish trust in the model, the user has to only reveal his information to the buyer after the transaction has been confirmed. Instead of asking the users to reveal the exact information attribute-value, we asked them to choose the data range that best reflects their information values that they would like to sell. Using these values we calculate the distribution as seen in Figures 4.3, 4.4, 4.5 and 4.6. These figures show us the probability distributions of the user data for different information types (i.e. information categories). The distributions though not revealing the exact information does give a clear picture about the nature of the information that the buyer could avail of in the information market. From this range value provided, we calculate the amount information value for every information type.

From the probability distributions, we calculate the information value for the information shared by each user using using Shannon's information formula.This is shown in Figure 4.7. This value depends on the type of information a user shares and the amount of obfuscation present in that information. Solely looking at the information values, we can see that the value of user 8 in Figure 4.7 is quite less. It can be deduced from this that the information of this user is not in the 'in demand' category type required by the buyers in the information market.

Finally, combining the demand score and the amount of information value for

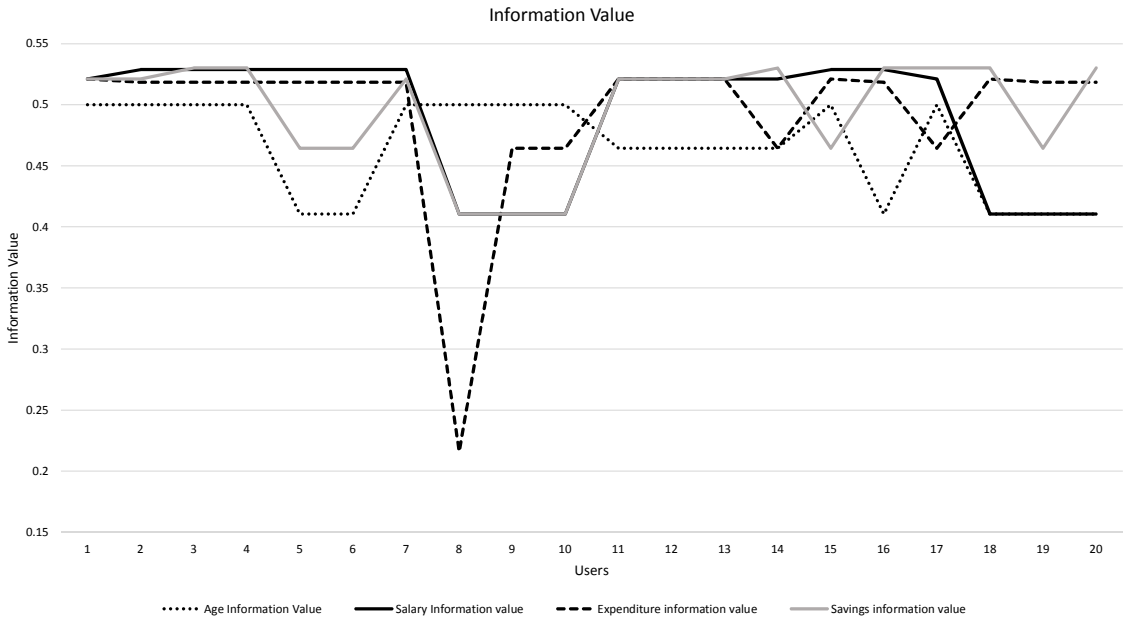


Figure 4.7: Value of information shared by users

each information category for each user, we calculated the price each user can get for the information a user chooses to share (shown in Figure 4.8). From the figure we can see that the price offered to user 8 is lesser as compared to the others. This can be attributed to the similar lesser information value of user 8's information as seen in the Figure 4.7. The information value tells us the amount of information that is revealed by the disclosure of the user's information. Thus more the information value, means more amount of information about a user is conveyed to the prospective buyer.

Using our model's preliminary results, we can thus see that the user's information price is adequately proportional to the corresponding demand in the information market for an information type and the information value for a user.

4.3 Pricing mechanism with exponential demand gauging

While we have Shannon's information theory to guide us to understand the amount of information that has been divulged, to maintain a fair balance in the information market, we need to incorporate the buyer's demand in the market also to decide on

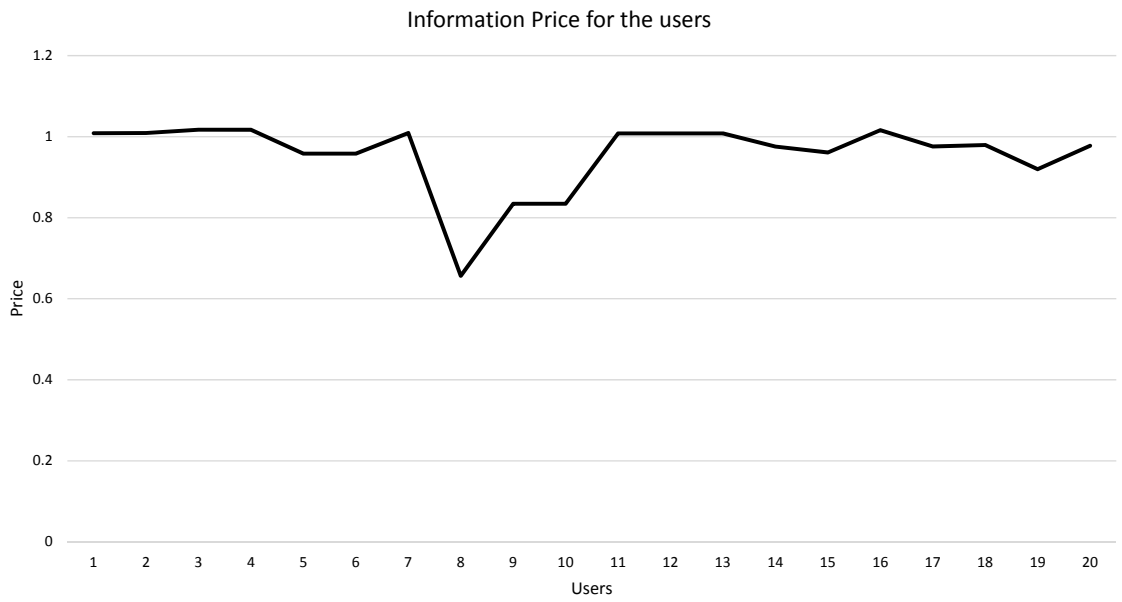


Figure 4.8: Price for each user‘s information

the price. In this section we have considered the scenario where the buyers could be dishonest and could collude with others and possibly state their willingness to pay untruthfully.

To gauge the demand from the buyers, we ask the buyers to state their willingness to pay, also called as demand. To be assured of the fairness of the prices the buyers state and to make sure that they are being truthful, we use the exponential mechanism [95] to keep the buyer's answers as close to their true valuations as possible. Thus buyers *'score'* the information categories or types giving an idea about the demand in the internet information market for the different information types.

Thus, every information category will have its own 'score'. The higher the score, it means the demand for that information type of category is high, and this means that, that information category or type should nett a higher price. Thus we can say that, $\boldsymbol{price \propto score}$.

Now each user is free to choose which information category attribute-value they wish to divulge. As mentioned in the discussion above, each user can get an idea about the amount of information divulged by each information attribute-value using Shannon's information theory. To understand the value of his information, instead of

asking the user to reveal the actual information, we ask the user to mention the range in which his information falls into. This information value that we have calculated states the amount of information that is conceded when the user reveals or shares his information. From the point of the user, this amount is the risk for the user in the information market when he or she will choose to sell his information since this also represents a loss of his privacy. For a continuous interval, as explained in the previous section, the probability distribution can be the range in which the values fall under. This applies to our scenario of asking users to enter the range of their information values. Thus, since the range is the interval for every information type or category, $\boldsymbol{(b-a)>1}$ which means that the risk won't get negative. This is the user's 'risk'.

Now, greater the value of *risk*, greater should be the compensation, i.e. the price, be given to the user. Thus we can say that $\boldsymbol{price \propto risk}$.

Combining the idea of *score* and *risk*, we can arrive at the fact that $\boldsymbol{price = risk * score}$.

In this manner, each user can calculate the price for each of his information attribute-value and obtain a revenue equal to the sum of the price for each information attribute-value. Hence, ***Revene, R,*** can be obtained by $\sum \boldsymbol{price_{ivalue}}$ where $\boldsymbol{ivalue}$ is each *information attribute-value that is shared.*

Pricing model with exponential demand

for each user **do**
  for each information category k **do**
    Calculate the information value $\boldsymbol{value_k = -p_k log_2(p_k)}$
    Obtain demand from buyers using exponential mechanism *expoDemand*
    Calculate $\boldsymbol{price_k = expoDemandk * value_k}$
  Calculate revenue for each user $\boldsymbol{revenue = \sum price_k}$

4.3.1 Results & Discussion

We utilized the "Online Retail" dataset from the UC Irvine Machine Learning Repository (http://archive.ics.uci.edu/ml/) to test our model on the data. The dataset had 541910 entries which we analyzed and eliminated entries which had discrepancies in their values or had absent values. From the remaining entries, we isolated the

following information categories:

- **Amount a customer purchases in one invoice** - This gives the information of the amount that each customer spends in an invoice. This shows recurring customers and one time customers. This type of information can give useful insights about the spending power of customers and can help organizations to understand how to convert one time customers into recurring customers and also how to price items the attractive range that makes maximum profit.
- **Total number of items of each stock a customer purchases** - This gives the information about the quantity of a particular item (indicated by the stock code) that each customer purchases. Customers could either purchase the same item multiple times or multiple customers could be purchasing the same item. This type of information gives organizations an understanding of the type of items that are popular and in-demand by customers. Knowing this, they can then produce similar items or sell similar items to attract more customers thus increasing their sales.

We categorized the data from the information categories into range bins in order to calculate the information value that is lost when the data is sold or shared with the prospective buyers; refer Fig 4.9 and Fig 4.10. From the figures we can see that the information value is higher for some range than the others (range 201-400 in Fig 4.9 and range 1001-2000 in Fig 4.10). This is because these ranges contain the most information from the maximum users and thus any information shared from this range has a higher information loss. Possessing information from this range can reveal the most about the users since the bulk of the users fall into this range. We can call these categories as the "high-risk" categories.

We solicited the help of 5 participants who took on the role of buyers and we asked them to rate their willingness to pay or the amount that they would pay for the above information categories. Out of this we then applied the exponential mechanism to determine the best price that would maximize a user's revenue; refer Fig 4.11 and Fig 4.12. The winning price, which was selected on the basis of the exponential mech-

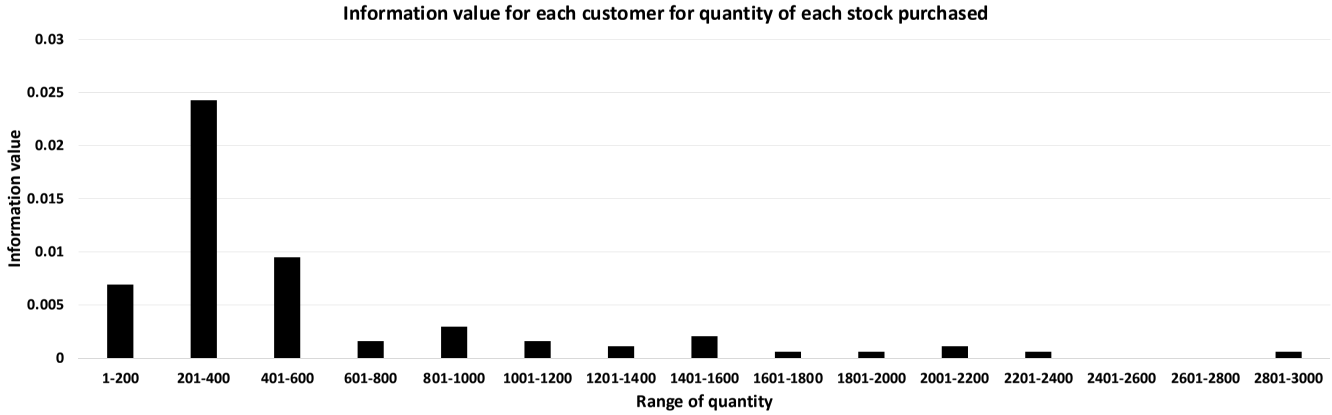


Figure 4.9: Value of information for customer for quantity of each stock purchased

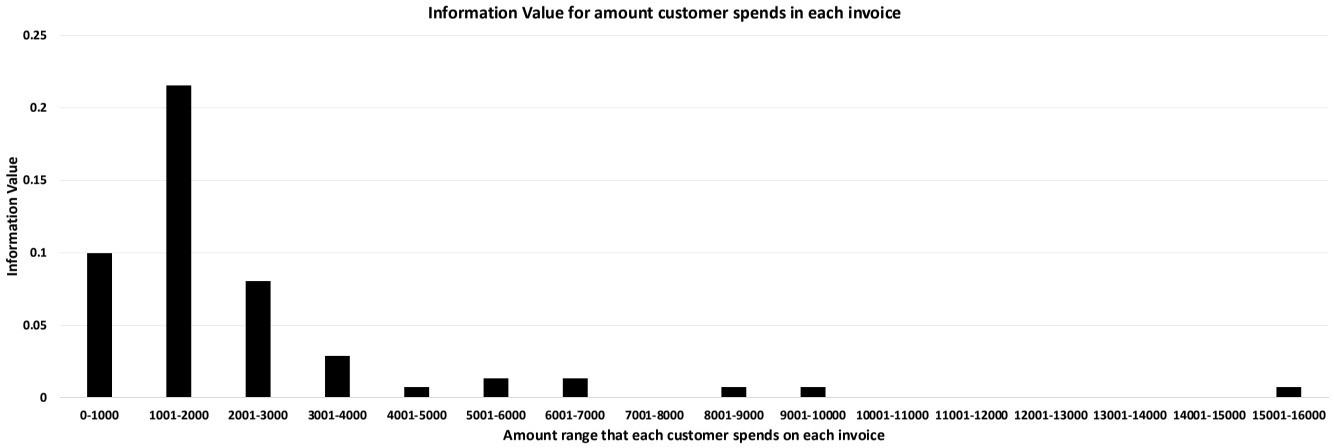


Figure 4.10: Value of information for customer for amount spent in each invoice

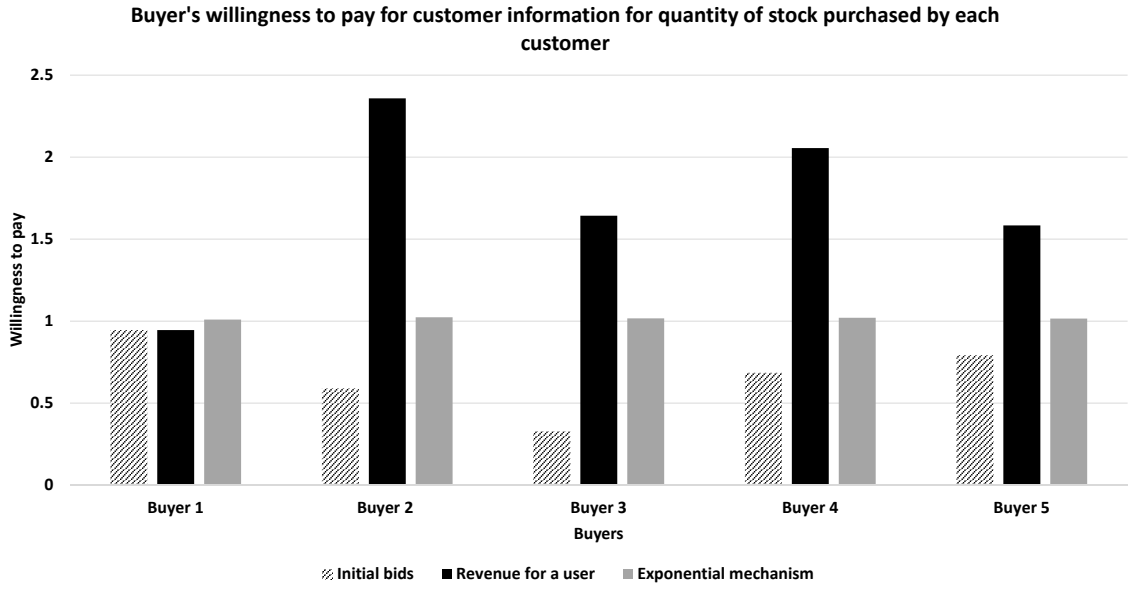


Figure 4.11: Buyer demand for customer information about quantity of stock purchased

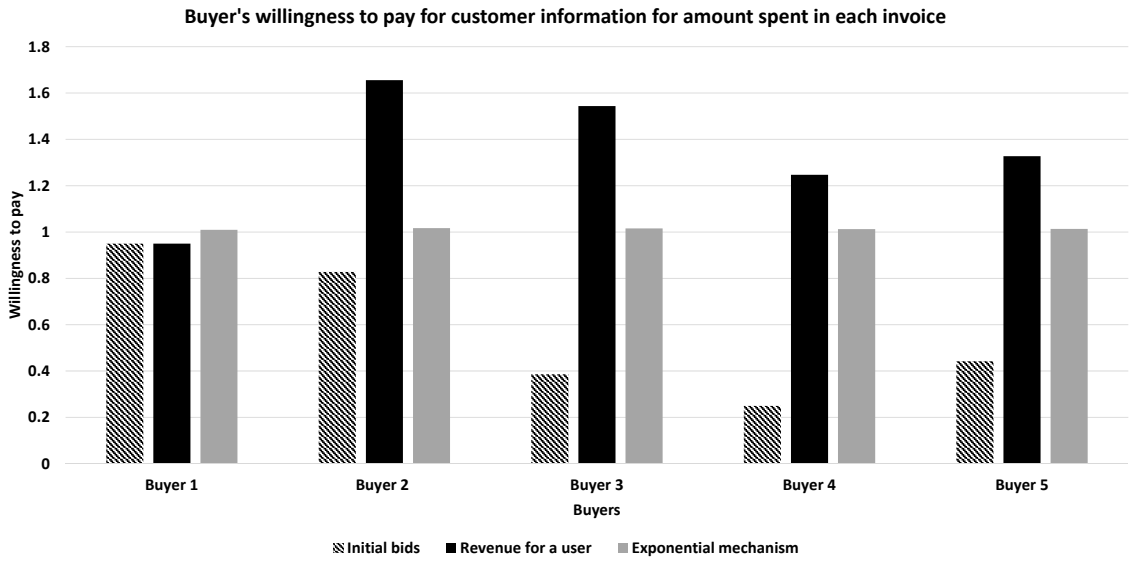


Figure 4.12: Buyer demand for customer information about amount spent on each invoice

anism, was then chosen to calculate information price for the customer's information for the different types of information; refer Fig 4.13 and Fig 4.14. We chose to ask them to rate their willingness to pay for the two information categories separately instead of choosing the same price for both was because different buyers are looking for different types of information for their benefits. Generalizing their willingness to pay across different categories would be unfair to the buyer.

We then consolidated buyer's price with the information value (which represents the risk for the user when he shares his information) to generate the information prices for the two information categories for each customer which can be seen in Fig 4.13 and Fig 4.14. As can be deduced from the graphs, certain users get superior prices for their information as compared to the other users. This is because these users have information in the "high-risk" categores and thus their monetary compensation for this information is higher as compared to the others. But on the whole each user stands to make some amount of money (greater than the current status quo where they make no money) from selling their information.

Combining the prices for both the information categories for each user we arrive at the total reveue for a customer which is seen in Fig 4.15. This is the total revenue of each customer from the sale of both his information categories to the buyer. The higher revenues seen in the graph correspond to the users who have information in the "high-risk" categories.

Though the type of information shown in the results are related to retail businesses, our model could be applied for any type of businesses or scenarios with an online presence (for instance for house agents to understand customers looking at houses). The types of information that can be valued can also be varied - from browser information, to the number of hours a day the user spends online to the educational and location information. Each of this information can be captured and sold to the prospective buyers to generate revenue for the user.

[111] also uses the exponential mechanism in the form of auctions to sell user information. Relying solely on auctions puts the control over prices entirely in the hands of the buyers without considering the value of the information that is lost. The information value represents the user's cost in the transaction that should be rightfully compensated. Our above method includes the valuation of the information that is shared with the users.

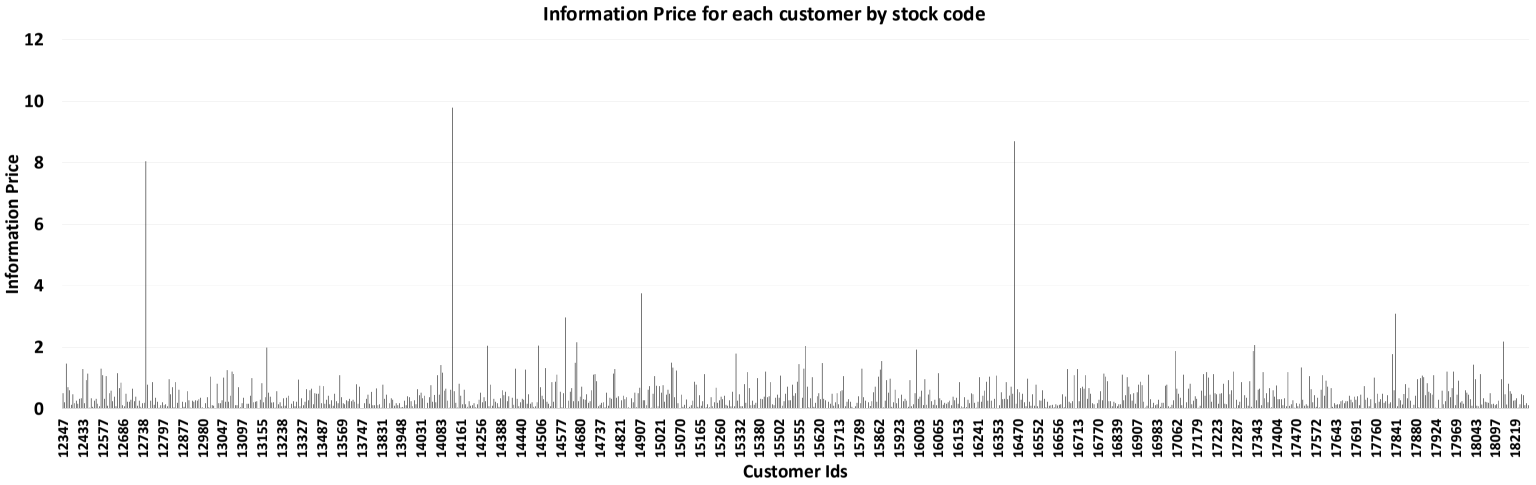


Figure 4.13: Information price for the quantity of stock purchased

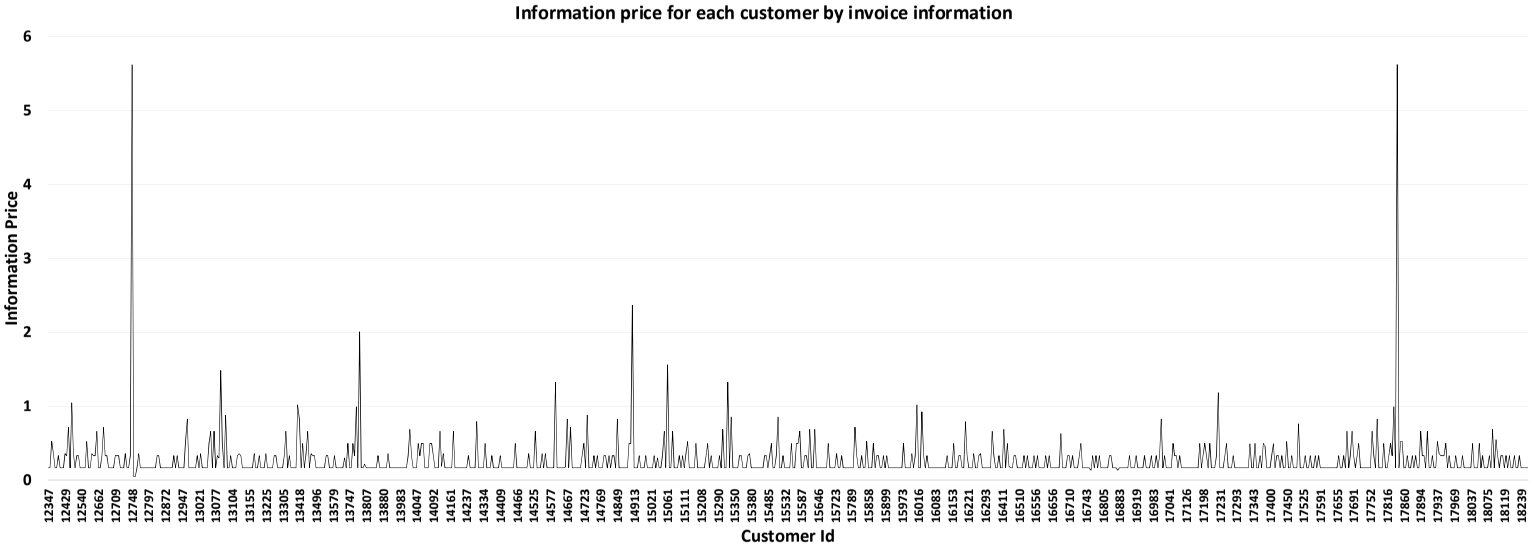


Figure 4.14: Information price for the amount spent in each invoice

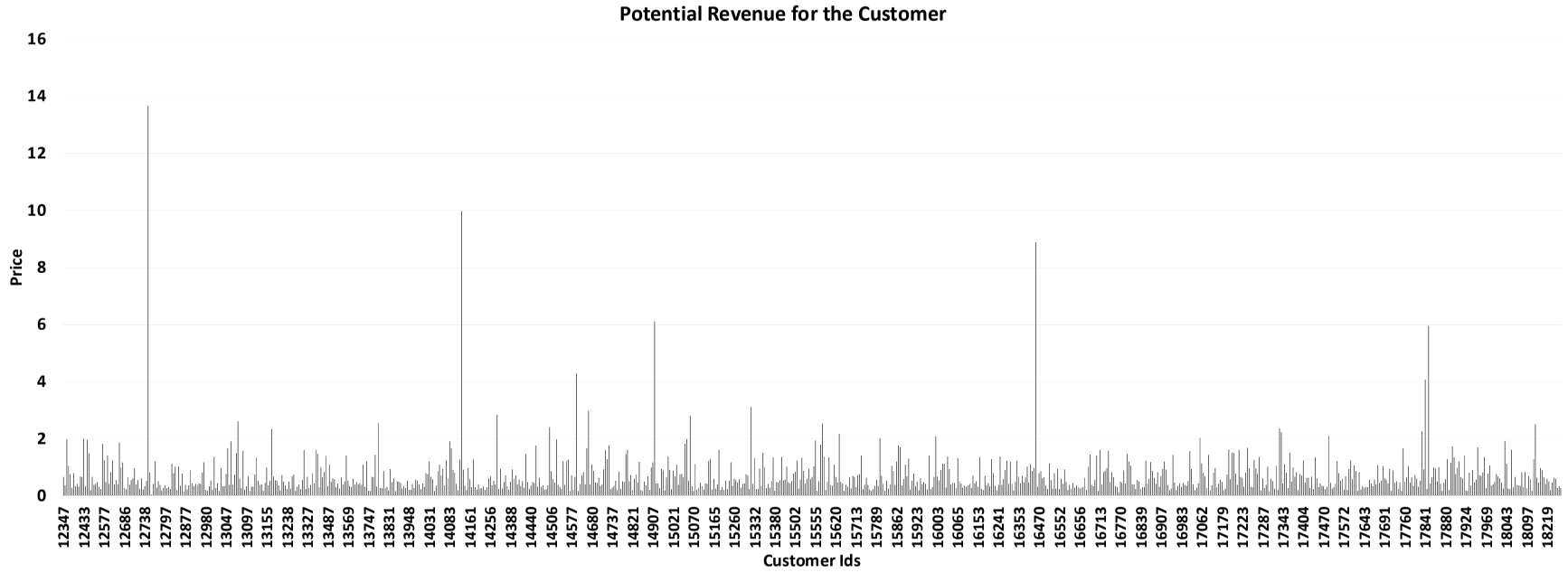


Figure 4.15: Potential revenue for a customer

4.4 Summary

For organizations, all the information about and generated by the users and the associated insights from this information, which make up it's big data, are it's biggest asset. Any and all information about the user is the meal ticket for the data brokers and organizations to earn a huge amount of money. And users just seem apathetic to this obvious infringement of their information and their privacy.

The value of an object is more easily understood to a layman when it is put in terms of a monetary value. In this chapter we have presented our pricing model that allows a user to realize the value of his information and the potential revenue he can make off of it.

In the scenario described above, we presented the information using the concept of Shannon's information theory to understand the value of information that is lost when it is revealed. We also presented the demand analysis using two methods - the regular method and the exponential mechanism method. In the regular method, there is a possibility of untruthfulness and collusion that may cause the pricing of information to not be calculated in the optimal way. To counter this, we have tested the the demand by applying the exponential mechanism which ensures that the buyers are truthful and honest in stating their demand. The exponential demand method provides the buyers with the least incentive to lie about their true valuations and thus enables a more accurate pricing for the information in question. The reflection of adopting the exponential mechanism on the user revenue is minuscule and gives the user more trust to participate in our pricing model.

5

Pricing with privacy - the utility comparison method

Looking at the enormous amounts of information being collected and collated today, big data has now become a pervasive term used in almost all organizations. With such importance being attributed to big data, it is no surprise that a lot of research has dealt mainly with harnessing the power of this big data to help organizations improve themselves. Amidst all this what has gone largely unnoticed is, exploring the 'value' of this big data, especially from the point of view of an internet user, from whom all this information is collected. Without this, these users have no idea about the value of their information and hence are not compensated economically for the same.

In this chapter, we have presented our pricing models keeping in mind the privacy

conscious user who would like to make money by selling his information, but also would like to maintain a semblance of privacy over his personal information.

5.1 Introduction

A revenue maximizing risk-neutral person who would like to sell information, can mask his information and sell it to a potential buyer if the seller is completely controlling the information [140]. [100] discusses the statistical overview about the concept of noise addition for the problem of balancing the utility of information while maintaining privacy by introducing noise to the data.

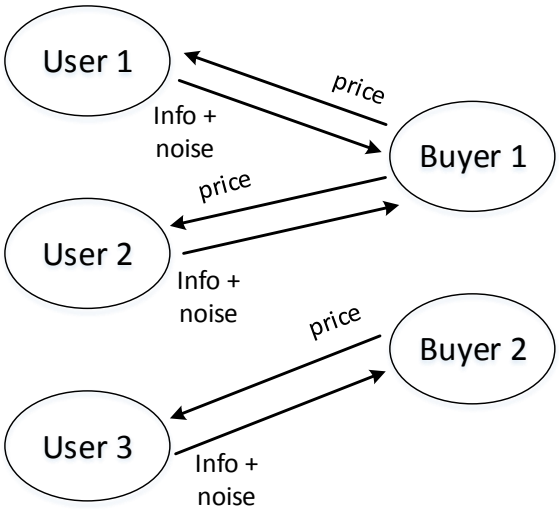


Figure 5.1: Scenario I - where a single user adds noise

In our scenario we have the user who wishes to profit from the sale of his information and we have the buyer who wishes to avail himself of this information for a price. But the user is worried about his privacy. To incorporate this concern, the buyer is willing to accept certain noise (or distortion) to the information but at a price.

We propose that the price the buyer is willing to pay is inversely proportional to the amount of noise in the information he is willing to tolerate. In other words $\boldsymbol{price \propto 1/noise}$ The following is an overview of our idea:

- The user states the type of information he or she is willing to sell restricting

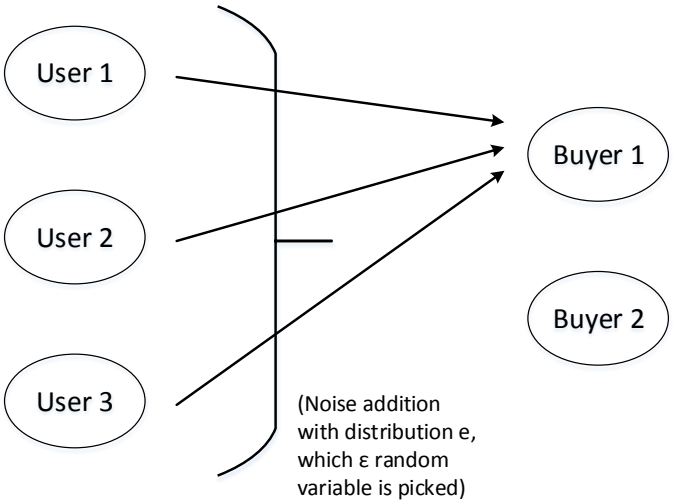


Figure 5.2: Scenario II - where users can form groups and collectively add noise

himself to information of numerical type for the purpose of our model (eg: salary, health data etc)

- The buyer chooses from this and states his willingness to pay or willingness to buy (allowing for the introduction of noise)
- Once this is known, the user can choose either of the following:
  - Scenario I: Choose the amount of noise to be added (from a randomly chosen distribution) with more noise for less price and vice versa as shown in Figure 5.1
  - Scenario II: Group with other users and decide on the maximum amount of noise to be introduced as shown in Figure 5.2
- Once this distortion is done, the user can give his or her information to the buyer assured of the privacy of their information.

The user(s) can then repeat this process with other buyers as desired.

In the following sections, we have expounded more about our *Scenario I* where the user adds obfuscation to his information. The buyer states his interest or demand for this information with the obfuscation and the pricing for this information is judged accordingly.

5.2 Pricing model

We believe that, data owners should have more control over their information especially with regards making their information private. And this protection of confidentiality often comes at a price of the information that one is trying to protect. This protection means that the data would need to either be perturbed or altered in some ways. There are many such statistical disclosure limitation strategies that through the use of various methods (like adding noise to numerical values, swapping data values etc) attempt to obfuscate information in order to protect the privacy of the user. But the implementation of these methods makes the data less useful for analysis purposes, i.e. the utility of the data decreases with an increase in the obfuscation intensity.

Why is data utility important? When a buyer buys a good, he needs to know the quality of the good. With goods having tangible properties, it tends to be easy. But with goods having intangible properties like information, it gets much harder. Especially if you also want to address the issue of user privacy. But if the utility of the information is known or rather if it is known by how much has the utility of the information been reduced or lost due to the introduction of obfuscation, then this information could be used for large scale data analysis without worrying too much about the reflection of this obfuscation on the results.

But the question that arises is how can a buyer judge the quality of the obfuscated information? If there is obfuscation, why would a buyer purchase this information especially without knowing how good this information is. To address this problem, of judging the quality of the information, there have been works that discuss evaluating the utility of the obfuscated information. This method of adding controlled noise to protect the privacy of the information and yet make it available for analysis is known as "Statistical Disclosure Control" [65] [70].

In this chapter we have incorporated the approach of comparing the Cumulative Distribution Functions (CDFs) to decide on the data utility as in [65] which mentions the *Kolmorgorov statistics*.

$$\boldsymbol{D(S_1, S_2) = sup_{1<j<n+m} |S_1(z_j) - S_2(z_j)|}$$

where, $\boldsymbol{S_1, S_2}$ are the CDFs of the original and obfuscated data respectively. $\boldsymbol{n, m}$ are the number of entries of the original and obfuscated data. $\boldsymbol{z_j}$ is the $\boldsymbol{j^{th}}$ entry of the pooled data.

The Cumulative Distributive Function gives as an output, the probability that a random variable would have a value that is either less than or equal to the argument of the CDF. In our approach we have leveraged the usage of the CDFs because it helps to entirely describe the probability function. In many methods, the using the CDF as a basis to measure the degree of differences in the distributions obtained from masked and original data have shown much success. The Kolmogorov statistic is used as a check to see how different two sets of comparable data are. It analyzes the relative distribution of the datasets and thus can be used as the basis of robust analysis.

Now if the utility of the data, i.e. $\boldsymbol{D(S_1, S_2)}$ is high, that means that the difference between the original values and the obfuscated values is large, which means that they may not be much useful. So, a smaller value of utility is what is valuable and should be priced more, i.e. $\boldsymbol{price \propto 1/utility}$.

The utility of the information cannot be the only factor that decides the price, the demand in the information market also decides the price. In this chapter, we have combined the utility of a information category along with the demand in the information market for that information category. The demand in the market shows us the interest among the buyers for a information category. To gauge the demand from the buyers, we ask the buyers to state their preference by rating the different types of information categories. Thus buyers *'score'* the information categories or types giving an idea about the demand in the internet information market for the different information types.

Thus, every information category will have its own 'score'. The higher the score, it means the demand for that information type of category is high, and this means that, that information category or type should nett a higher price. Thus we can say that, $\boldsymbol{price \propto score}$.

Combining the idea of *score* and *utility*, we can arrive at the fact that $\boldsymbol{price =}$

$\boldsymbol{score/utility}$.

In this manner, each user can calculate the price for each of his information category and obtain a revenue equal to the sum of the price for each information category. Hence, ***Revene, R,*** can be obtained by $\sum \boldsymbol{price_{infoCategory}}$ where $\boldsymbol{infoCategory}$ is each information category that is shared.

In our research which focuses on pricing information, we believe that the concerned stakeholders (the buyers and sellers) must all have their needs balanced and incorporated. This is the reason we have incorporated the idea of privacy and data utility and the information demand in the information market to generate our model for information pricing.

Pricing model with obfuscated user data

---

for each user **do**
  for each information category k **do**
    Obtain CDFs of original and obfuscated user data
    $\boldsymbol{S_1} \leftarrow$ CDF of original data
    $\boldsymbol{S_2} \leftarrow$ CDF of obfuscated data
    Calculate $\boldsymbol{D(S_1, S_2) = sup_{1<j<n+m}|S_1(z_j) - S_2(z_j)|}$
    Obtain buyer's demand: *score*
    Calculate $\boldsymbol{price = score/utility}$
  Calculate revenue for each user $\boldsymbol{revenue = \sum price_k}$

5.3 Experiments

We utilized the "Online Retail" dataset from the UC Irvine Machine Learning Repository (http://archive.ics.uci.edu/ml/) to test our model on the data. In the real world, we would be incorporating data from the users and then collating it to deliver to the buyers.

The dataset had 541910 entries which we analyzed and eliminated entries which had discrepancies in their values or had absent values. From the remaining entries, we organized them into the following information categories:

- Amount a customer purchases in one invoice
- Total number items of each stock a customer purchases

- Total amount a customer spends

We then plotted the CDF of these information categories. To this data, we then introduced random noise with the distribution similar to the existing information categories and then plotted the CDFs of these values. (See fig 5.3,5.4,5.5)

In the next step, we calculated the difference between the CDFs (i.e. the difference between the CDFs of the data with and without the noise), which gave us the utility of the data values. We then selected the maximum of these utility values for each of the three information categories.

To gauge the demand in the market for the information, we solicited the help of 10 participants to take on the roles of buyers and explained to them our model and told the participants to state their interest in a particular information category (as mentioned above) that they were most interested in.

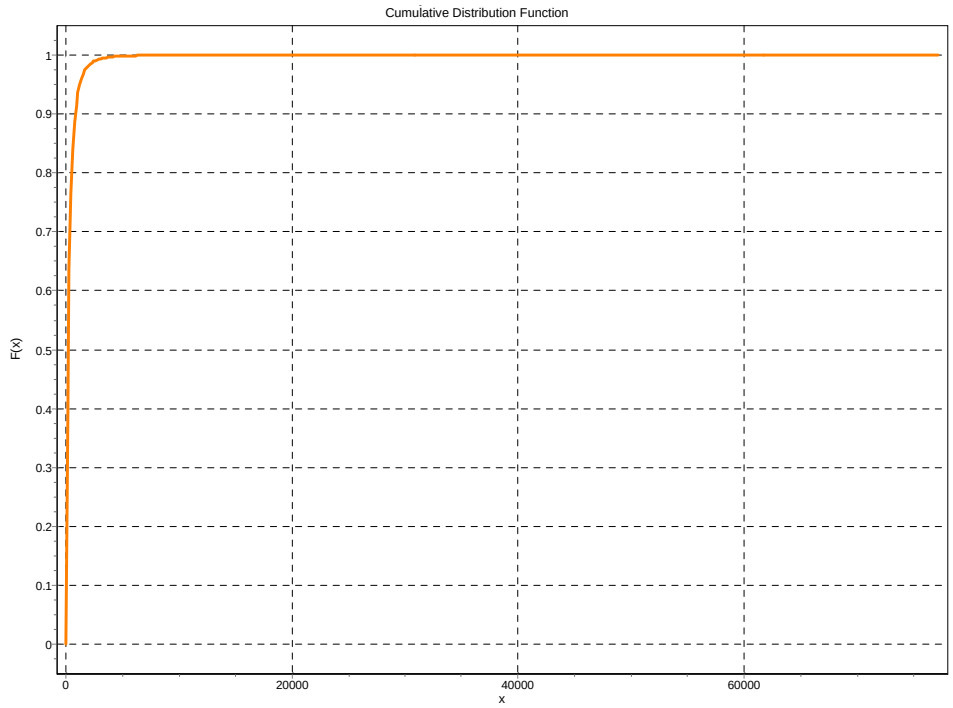
After meticulously charting out the demand, we combined the demand and the utility to determine the price for that information category.

5.4 Results & Discussion

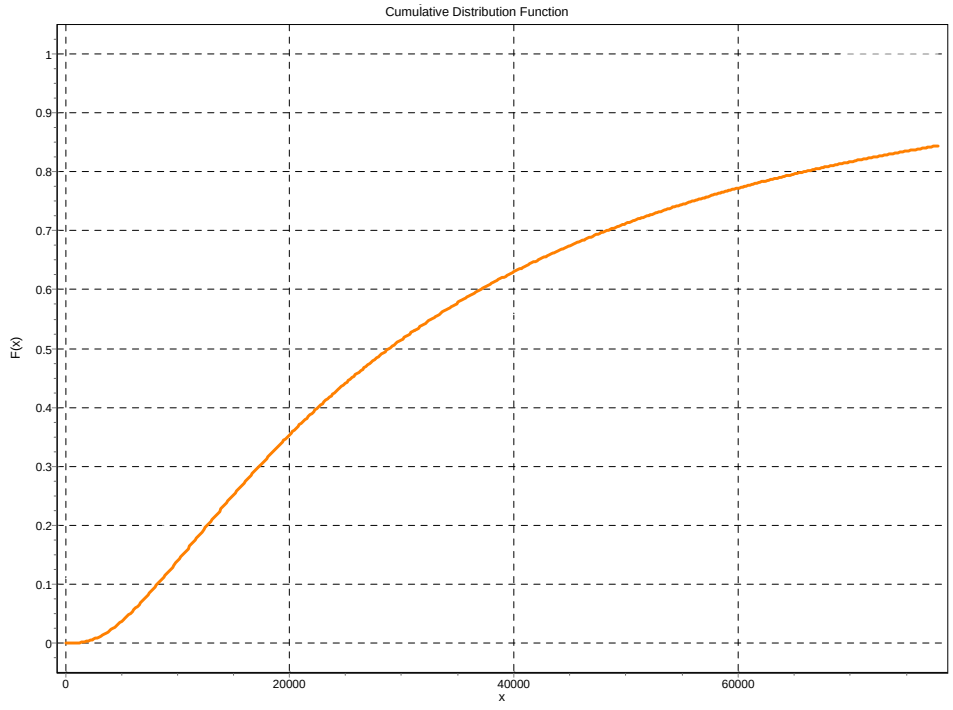
The results of the utility comparison and demand analysis gave us the results as seen in Fig 5.6. From the figure, we can see that the demands for the three information categories are in the similar range. But the utility of the information category *"Total amount a customer spends"* is significantly higher than the utilities of the other two information categories. This is why the price for this category is lower than the other two despite the high demand.

The price of the information category *"Amount a customer purchases in one invoice"* is higher than the other two information categories. This is because the utility is lower for this information category and the demand for this information category is higher in the information market.

This shows us that there are two factors - *utility of the information category* and *demand for that information category* that decide the price of the information. Both these factors work in unison to balance the pricing needs of the sellers of the

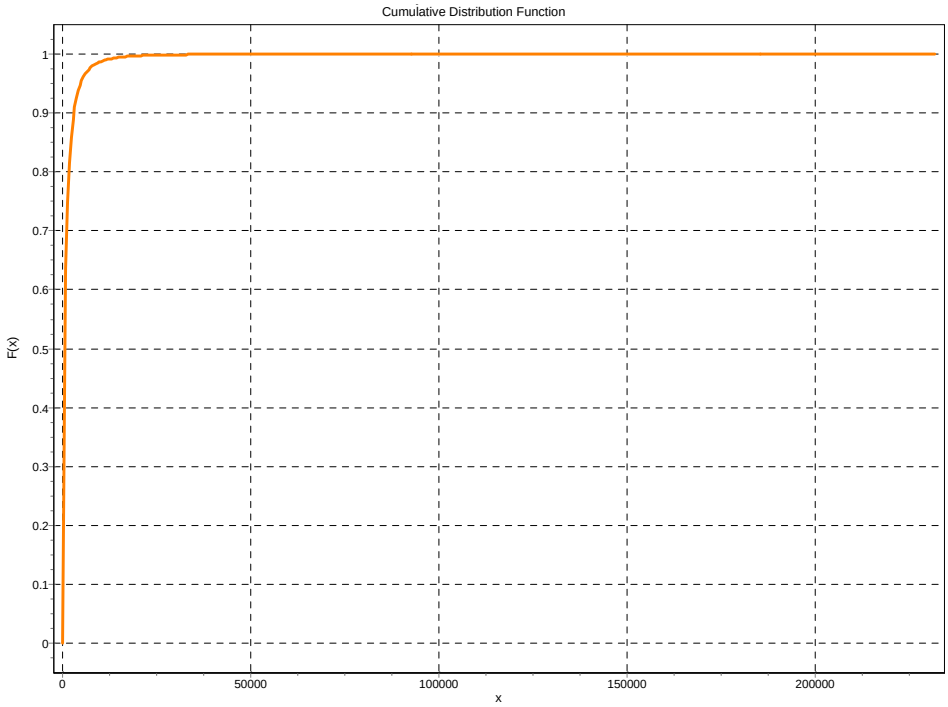


(a) without noise

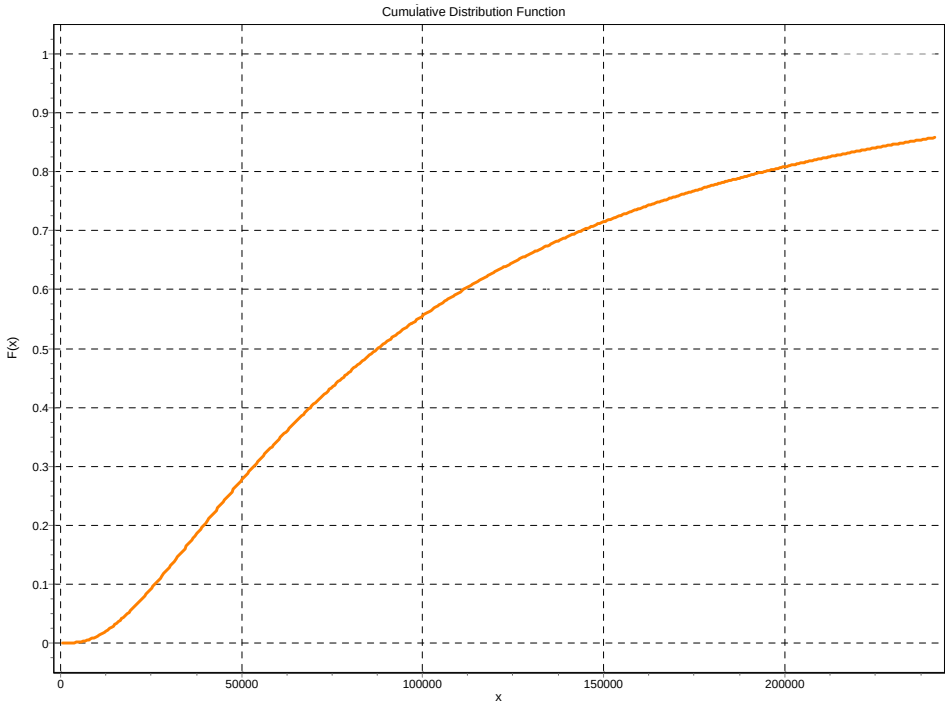


(b) with noise

Figure 5.3: Amount a customer purchases in one invoice

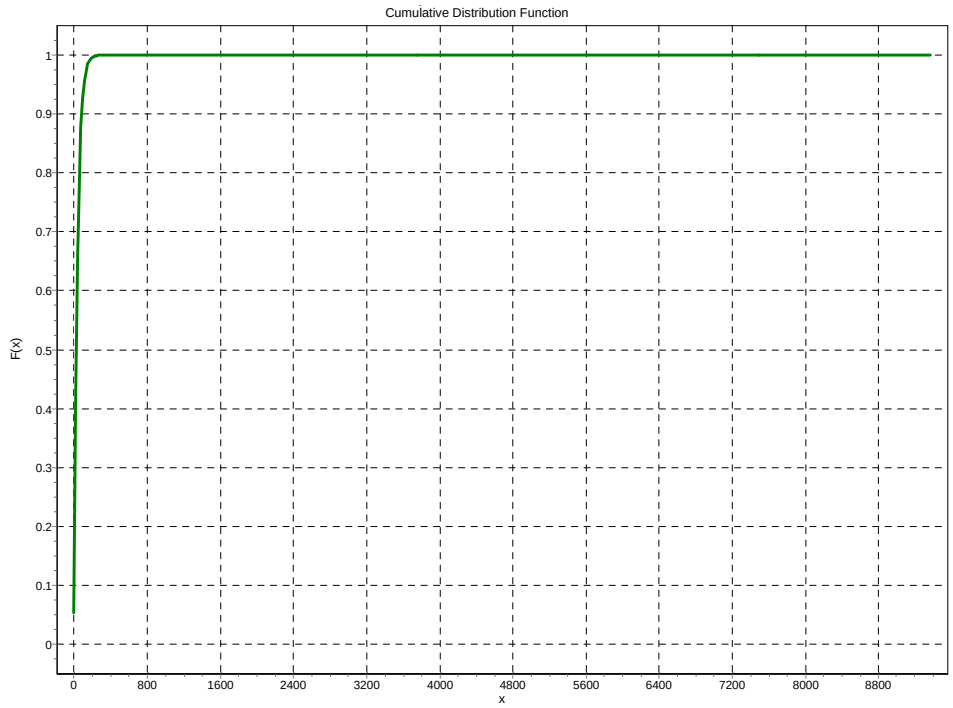


(a) without noise

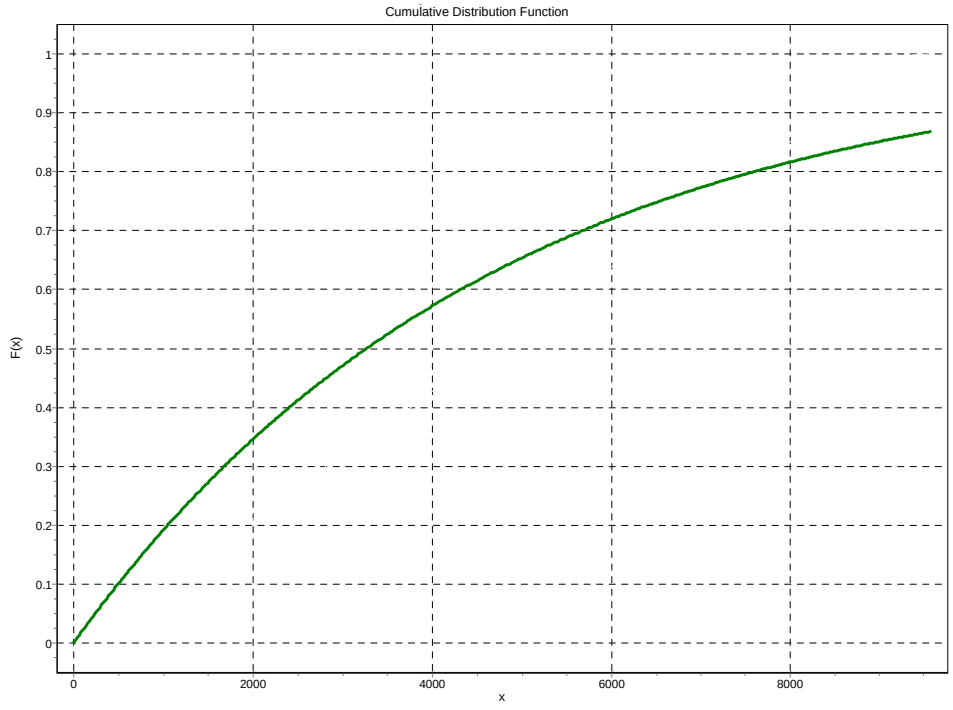


(b) with noise

Figure 5.4: Total amount a customer spends



(a) without noise



(b) with noise

Figure 5.5: Total number items of each stock a customer purchases

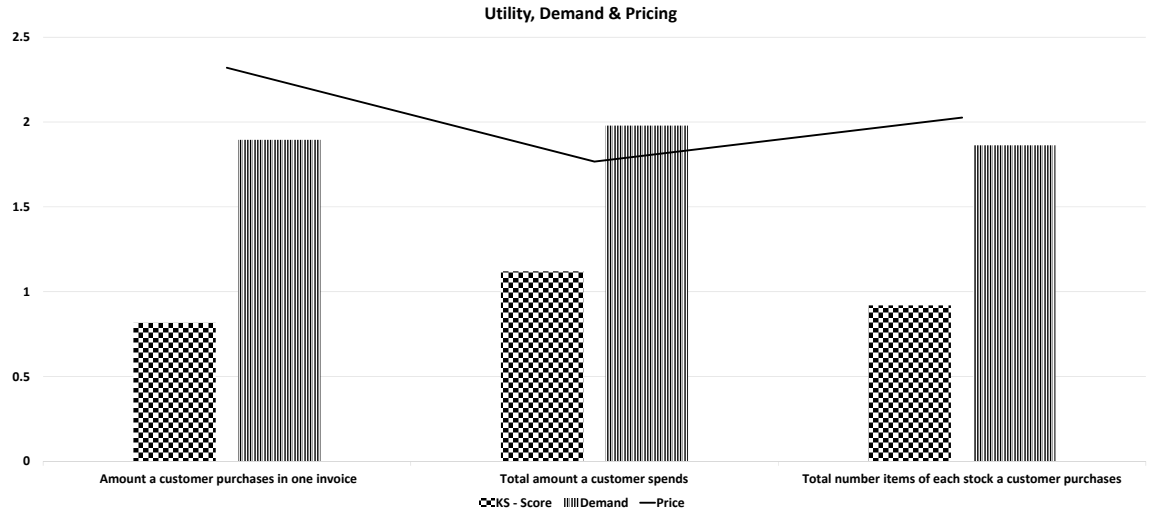


Figure 5.6: Graph showing utility, demand and pricing of the three information categories

information and the buyers of the information.

The buyers can look at the utility score and judge the quality of the information themselves and at the same time the users are assured of the privacy of their information.

5.5 Summary

In this chapter, we have provided for the user‘s concerns about privacy and at the same time provided for a means for the user to sell this information. This method provides a fair and balanced mechanism to tackle the question of information pricing in a simplistic and practical manner.

The scenario in this chapter targets numerical information. The problem of protecting privacy while sharing other types of information categories is a possible avenue for further research. Our method allows users to judge for themselves the demand for different types of information and decide for themselves which information they want to reveal. Our pricing model allows for user privacy by obfuscating the user information and adjusts the pricing accordingly. Our model does not in any way depend on the number of the users or buyers and thus can be scaled as desired. Using our model eliminates the need to use any third party and connects the buyers and the users (the sellers of information) directly to each other.

6

Pricing with privacy - the risk comparison method

In the previous chapter, we introduced obfuscation in the user information and calculated the utility of the information using the distribution functions. In this chapter, we again introduce obfuscation to the user information. But we judge the quality of the information for the buyer after obfuscation by using the *Sharpe measure* and determine the price accordingly.

6.1 Introduction

The introduction of the noise in the information assuages the user that despite the sharing of his information, his privacy is still retained. For the buyer on the other hand, the obfuscated information is a risk that he is taking for the information that

he wants. The buyer would only be willing to pay for the information according to a risk adjusted measure. The pricing model discussed in the further sections describes the amalgamation of these two concepts.

6.2 Pricing Model

Each user possess information belonging to a certain category or type, eg. salary, age, location etc. And each type of information type or category has a particular demand in the information market. This demand shows us the relative importance of each information type for the buyers in the information market.

Now, the user is allowed to introduce noise in each of his information type. This distortion in the information then is the *risk* that the buyer must endure in order to obtain that particular information type. To enable the buyer to make a decision of his choosing, we must provide the buyer with a means to compare the available information types of the participating users along with the risk so that the user can purchase, i.e. *invest* in a particular information type of a user or group of users. To calculate this, we borrow from the financial world the concept of 'Sharpe ratio'.

The parameters to calculate the Sharpe ratio are:

- Average return from an investment 'A' (with risk): $\boldsymbol{R_A}$
- The risk free return from an investment: $\boldsymbol{R_f}$
- Standard deviation of the investment: $\boldsymbol{SD_A}$

Thus the *Sharpe ratio* can be calculated as:

$$\boldsymbol{(R_A - R_f)/SD_A}$$

Sharpe ratio can be either positive or negative. A negative value of the Sharpe ratio indicates that the risk free asset outperforms the risk adjusted asset in consideration.

In our model, we have a set of users who each have information from different information categories. The users would like to sell this information in the information market for a fair price. In order to ensure privacy, we allow users to introduce obfuscation in the form of noise that they add to their information from a normal distribution with varying variances for the different information categories.

We also have a set of buyers who would like to purchase this information. These buyers are willing to accept the distorted information but would like to understand the risk that they are undertaking by purchasing this noisy information from the users.

To calculate the 'value' of information, we utilize Shannon's information theory. The concept of Shannon entropy [117] has been used widely in communication system studies to understand the privacy and anonymity problems that arise. It is used mainly to calculate the amount of uncertainty for an outside bystander when it comes to figuring out role assignment (sender or receiver) in communication systems. Shannon entropy is used to assess the extent to which an outside bystander or observer would need additional information to classify the roles assigned in a communication system. In [24], the authors use Shannon information theory along with Option pricing theory to value privacy. They use it to calculate the value of an attribute of a user's information in their model. But they look at it from the point of view of option pricing and use it to determine the amount of privacy lost for a user.

In our model, we utilize the concept from Shannon's idea of amount of information gained by understanding it as the amount of information gained when an attribute-value from an information category is revealed. As explained in [86], Shannon's idea is to understand the number of questions that needed to be asked to correctly guess the attribute-value. When we have a case with discrete probability, $\boldsymbol{H}$, which is the term used by Shannon to indicate the amount of information, is equal to $\boldsymbol{H = -\sum p_i log_2(p_i)}$ where $\boldsymbol{p_i}$ is the probability of a particular attribute-value.

For the information market, the information value for the user's information without the noise is the 'risk free return' and the information of the user's information with the noise is the 'average return'. The variance of the normal distribution from which noise is added can be used to calculate the standard deviation (The square root of variance is the standard deviation).

For the purposes of our model, a risk free asset is information without obfuscation while risk adjusted asset means that the information has obfuscation. So a negative

value of the Sharpe ratio implies that the risk taken on by the buyer is more because the noise overpowers the original information, while a positive Sharpe ratio means that the risk free asset performs worse than the risk adjusted asset.

Pricing model by measuring risk

for each user **do**
  for each information category k **do**
    Obtain the risk free return $\boldsymbol{R_f}$ and Standard Deviation $\boldsymbol{SD_A}$
    $\boldsymbol{R_f} = -\sum \boldsymbol{p_i log_2(p_i)}$
    $\boldsymbol{SD_A} \leftarrow$ variance of the distribution of obfuscation measure
    Sharpe Ratio: $\boldsymbol{SR} = \boldsymbol{(R_A - R_f)/SD_A}$
    Obtain demand from buyers *demand*
    Calculate $\boldsymbol{price} = \boldsymbol{demand/SR}$
  Calculate revenue for each user $\boldsymbol{revenue} = \sum \boldsymbol{price_k}$

6.3 Experiments

To test our model, we conducted an experiment where we solicited the services of 20 participants who took on the role of the sellers i.e. the internet users who would like to sell their information. We asked each of them to enter the value for the following information types:

- Age (scores or marks)
- Monthly Salary
- Monthly expenditure
- Monthly savings

We then calculated the probability of the values of each information type, $\boldsymbol{p_{iType}(value)}$ using which we calculated the information value for each user using the formula $\boldsymbol{H_{infoType} = p_{iType}(value) log_2(p_{iType}(value))}$

We then asked the users to insert random noise from a normal distribution with a given variance and then calculated the information value from the noisy samples.

Now armed with the information value of the actual values (risk free) and the noisy values (average return), we calculated the Sharpe ratio for each user for each information type.

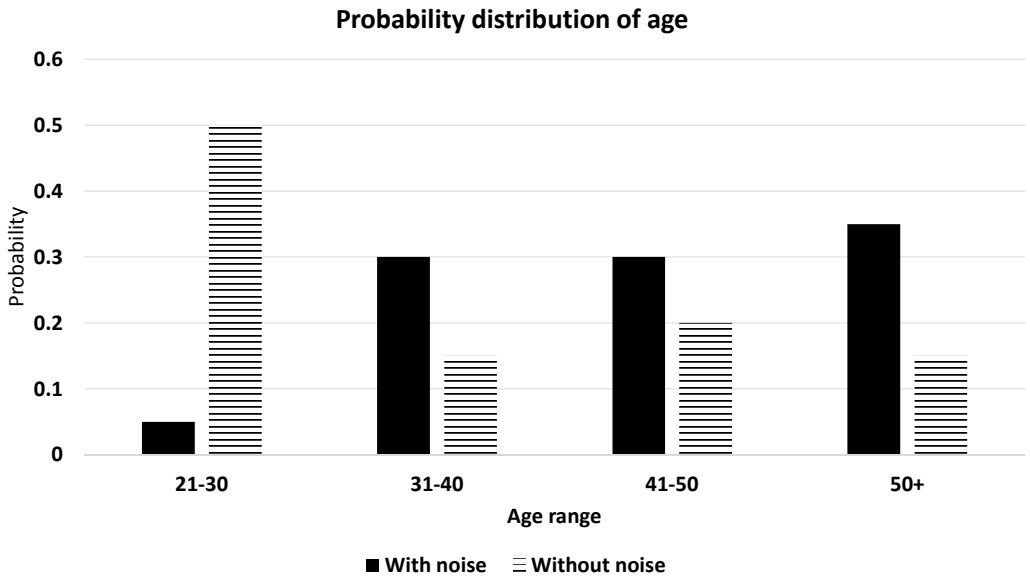


Figure 6.1: Probability distribution before and after introducing noise in age data

6.4 Results & Discussion

The Figures 6.1, 6.2, 6.3 and 6.4 show the probability distribution of the users based on their actual data and the probability distribution of the users based on their distorted data. As can be seen from the figures, due to the introduction of the noise, the actual values change causing a significant change in the distribution in the population. The buyer though isn't aware of the amount of actual distortion and can only judge the same based on the Sharpe ratio calculated.

Figure 6.5 further shows the information value calculated of the actual values and Figure 6.6 shows the information value of the distorted values. Judging from the figures, the distinctive changes in the information value are indicative of the change caused by the noise in the data. Solely looking at the information values, we can see that the value of user 10 in Figure 6.6 is quite less. It can be deduced from this that probably the information of this user is not found commonly in the population. Now it is upto the buyer if he wishes to invest in this information type of this particular user.

Figure 6.7 shows the Sharpe Ratio according to the various information types. The Sharpe ratio values of the information types salary and expenditure are in the

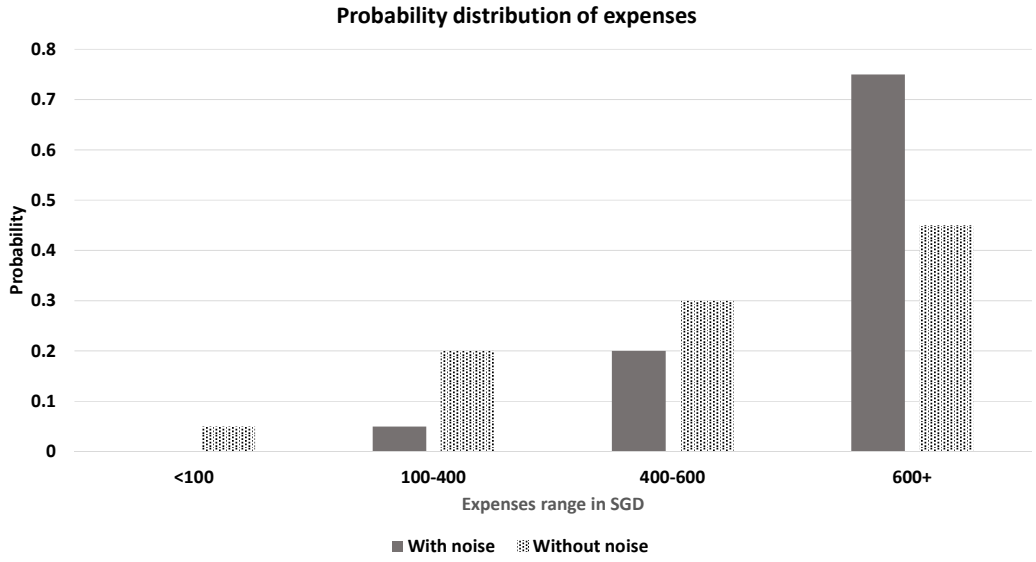


Figure 6.2: Probability distribution before and after introducing noise in expenses data

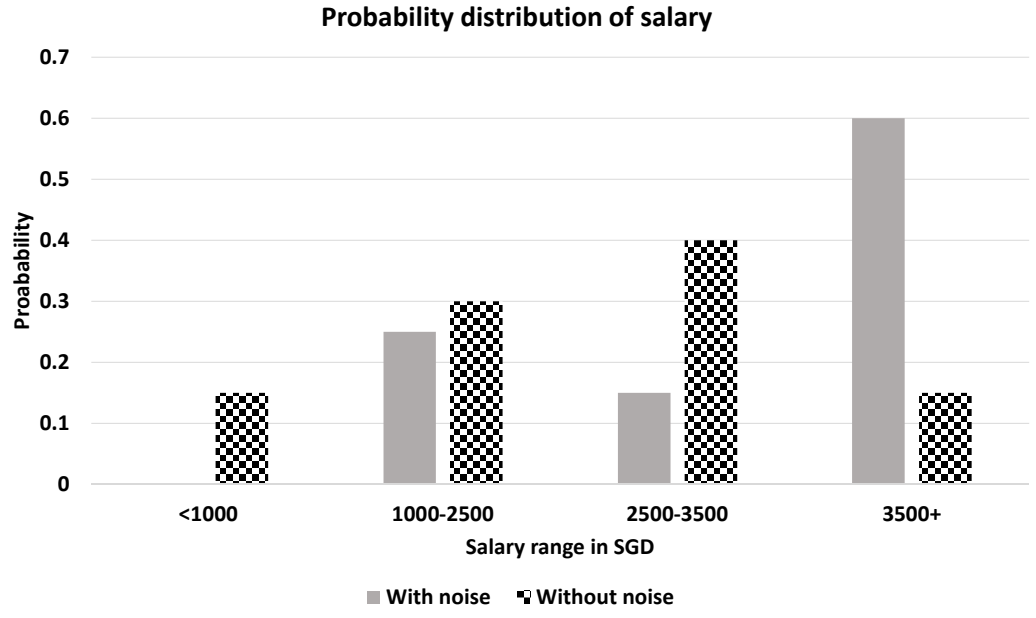


Figure 6.3: Probability distribution before and after introducing noise in salary data

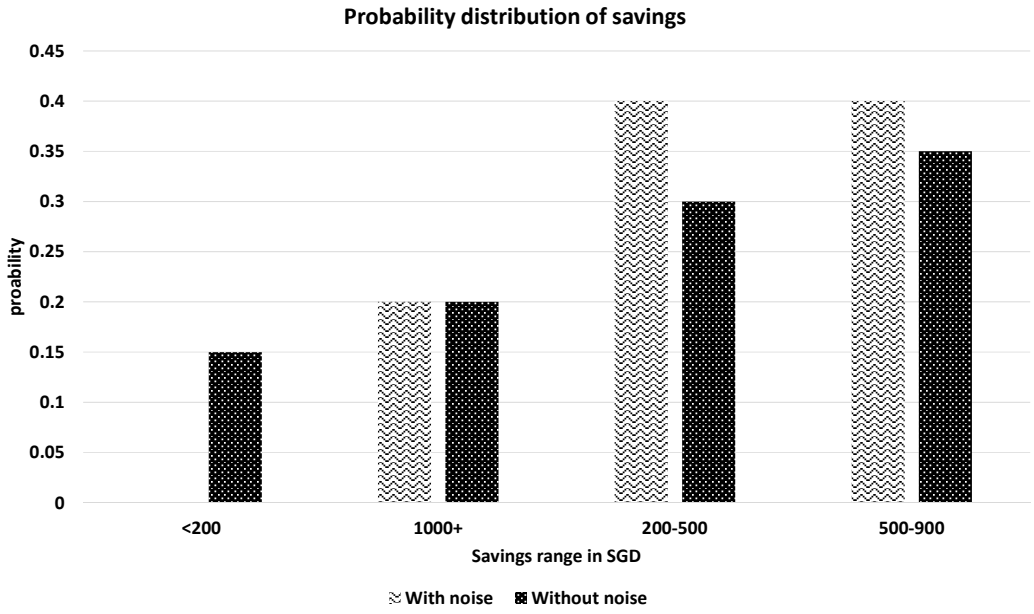


Figure 6.4: Probability distribution before and after introducing noise in savings data

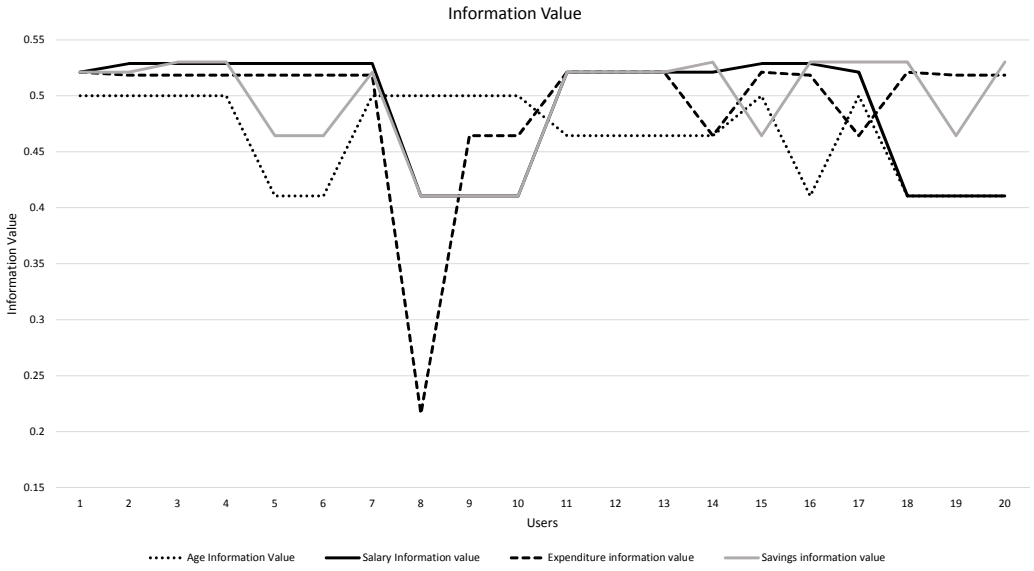


Figure 6.5: Information value of the actual data from users

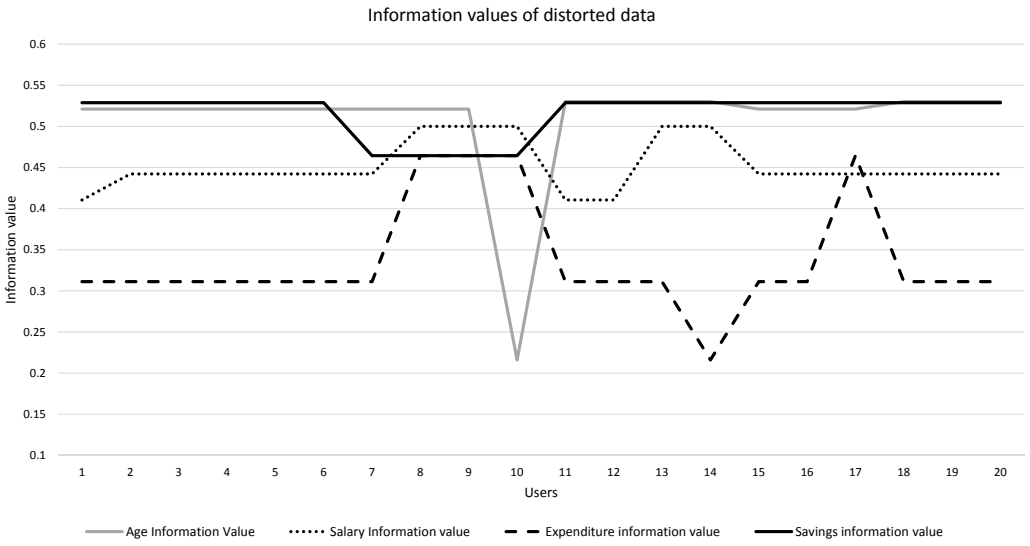


Figure 6.6: Information value of the distorted data

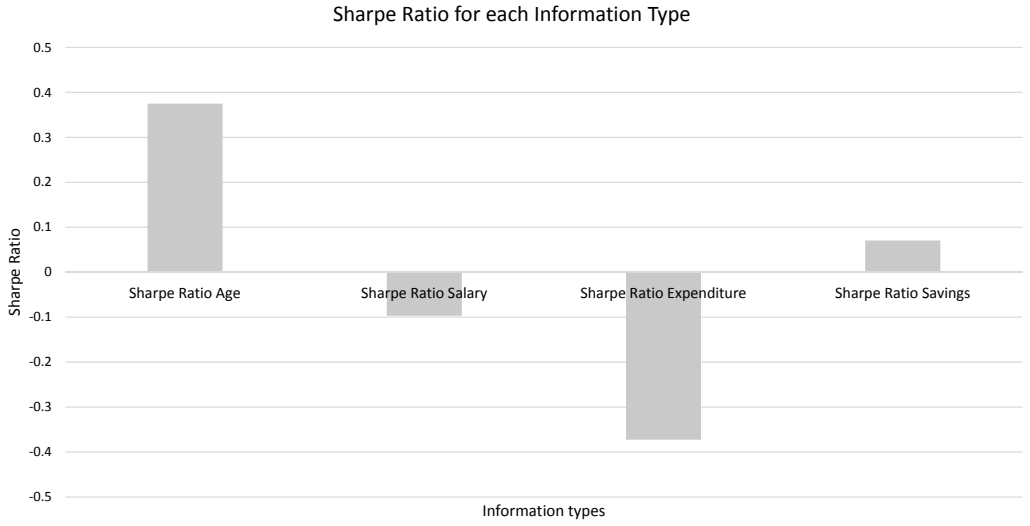


Figure 6.7: Sharpe ratio calculation by information type - age, salary, expenditure and savings respectively

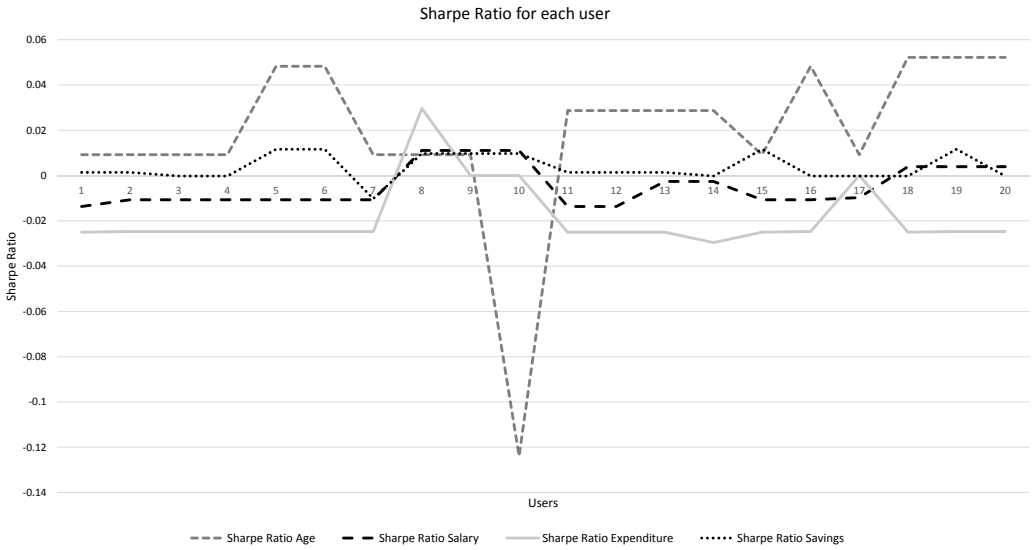


Figure 6.8: Sharpe ratio calculation for each user

negative range. This shows that the perceived risk(i.e. noise or distortion to the data) in these types is high as compared to the other two types. For certain calculations, the buyer is willing to digest this high risk and can choose to purchase these information types.

Figure 6.8 displays the Sharpe ratio for each user's information type. From the figure, it can be seen that certain users have negative values for their Sharpe ratio, especially user 10. In such cases, the buyer realizes that the perceived risk (or rather noise or distortion to the data) for the users with negative Sharpe ratios is on the higher end. Thus the buyer may not choose to purchase data from such users.

The users also, when made aware of their negative values of Sharpe ratio can choose to change the amount of distortion they introduce, thus making their Sharpe ratio more attractive.

Our preliminary results have helped us achieve the following:

- Given us a practical technique to make users aware of the value of their information in a population of their peers
- Allow the buyers to decide for themselves, based on the Sharpe ratio calculations, what information type he would like to purchase

The presence of this type of transparent mechanism in the information market is incentive enough to make both the buyers and sellers participate in the information pricing process together.

6.5 Summary

Our method is directly related to the information each user wants to sell in the online information market place. In this chapter, we have treated user information and the big data generated from it as an asset in the Internet information market.

To assuage the user of the privacy of his information, we have incorporated the obfuscation of the user information by the introduction of noise taken from a normal distribution. Now for the buyer, this obfuscated information is a risk for the buyer because the buyer is getting obfuscated information. This risk calculation is done by applying the concept of the Sharpe ratio. The Sharpe ratio is used to give a risk adjusted measure that can help the buyer calculate the return that they would get when they purchase this obfuscated information.The Sharpe ratio thus provides a fair measure for the buyer to assess the quality of information and decide on the prices for the obfuscated information categories.

The above described method caters to the concerns of the stakeholders involved in a fair and balanced fashion. The usage of the Sharpe ratio as a risk adjusted measure allows the buyer to know the average return he can expect from the information he wishes to purchase while at the same time the user being assured of a certain amount of privacy can sell his information to the prospective buyers without any hesitancy or misgivings.

7

Pricing mechanism modeled on Markov process

In this chapter, we present a model that prices information more from a buyer's point of view for a privacy insensitive user. This is for those situations when the prospective buyers of information would like to avail themselves of the actual information without it being obscured by any kind of obfuscation or noise.

Big data is the insight that is gleaned about a user from not just information about a user but also from information generated by a user's actions. All this insight is extremely valuable for organizations to turn non customers into potential customers or one time customers into returning customers. This also helps them to retain their customer base by catering to their personal tastes. Currently, users are not aware

that their information is being surreptiously monitored or analyzed for this purpose. So the users only avail the social benefit of having a personalized service but cannot avail of the monetary benefits enjoyed by the organizations.

To address this issue, we have proposed a model where the users can sell their information in the form of information bundles (explained in detail further in the chapter) to prospective buyers, thus availing themselves of monetary benefits off their information. We have chosen to model this as a Markov decision process.

7.1 Introduction

The markov decision process is a discrete time stochastic modeling process that is used for decision making situations [22]. It consists of a set of states, a set of possible actions that can be taken, a reward that is linked to the action and state and a set of probabilities that specify the probability of transition from one state to another based on an action. The markov model has been used for a variety of scenarios like for the "bidding decision making problem" [124] for the electric supply market, for the electricity demand response in the electicity market [129] and for the acoustic problem for dialogue systems [78]. The markov model is opined to be a scalable and surable model.

The current state in a Markov process characterizes the process under scrutiny. The state is responsible for acquiring the needed information from the previous states. Using the information from the previous states, the next action state can be determined. The effect that a particular action may have on the current state is reflected on only in that state. The idea behind this formulation is to come up with an approach to be taken when in a particular state that would promt the next action to be taken.

7.2 Revenue Generation for the user in the Internet information market

The scenario under consideration for this chapter is where the buyers would like to purchase user information directly from the user. This is to avoid the presence of any kind of "middle man" or "third party" in the transaction who could inflate the price that the buyer would have to pay. The buyer would also gain unfettered access to user information without having any kind of obfuscations or masking made to the actual user information. The buyer also has the freedom of choosing the number of users he would like to purchase this information from instead of the current scenario which has buyers purchasing information in predefined number of users.

In our scenario we have a set of users who would sell their information or sell access to their information given the right price and at the other end we have a set of buyers who are interested in purchasing this information. The users possess information in the form of 'information bundles'

Information bundles are sets of information (personal and/or private information of the user and could be supplemented with insights based on that personal information) which are clubbed together under some common category or type - Eg: '*Influence circle*' which could contain a person's contact list, most active contacts etc. Another category could be '*Places*' which could contain a list of the places a user has been to, the list of people who on the user's contact list who frequent similar places etc. Along with the information bundles, each user would have has his or her own private valuation of his information bundles which he or she may not like to share in the public arena of the information market.

We can model this as a Markov decision process because just like in the process, we are faced with a continuous arrival of the users and buyers into our model. We have not set any start and end points for the model, meaning that buyers and sellers are free to jump in as they please to engage in a monetary transaction. There is a certain amount of uncertainty involved with the participation of the users and the buyers in

the model. If the users are not happy with the price (their private valuations exceed what they are being offered) or are not interested in selling their information, they may not participate in the model. The buyers could also choose to not participate in the model because they may get a better offer for user information from other sources. Thus, the decision by the users and the buyers to participate in this process cannot be predetermined and can often be thought of as random. The conditional probability of generating the revenue depends on the decision of the user and the buyer to willingly engage in the current scenario. The action of the user to choose to sell his information results in him getting rewarded in the form of the revenue amount that the buyer would be paying the user for his information. Also, the decision of the stakeholders to participate or not participate does not have any bearing in the participation of the other stakeholders. This is why the technique of choosing the Markovian approach would help us determine the best possible revenue generation model for our scenario.

For traditional goods the demand can be calculated by analyzing previous prices, the different quantities of the good that were sold or the prices of the traditional goods can be experimented with to estimate the corresponding changes in the market. But for a non - traditional good like information, the demand analysis cannot be performed using the above methods because there is no precedence for pricing user information in a fair and transparent fashion.

The objective of the users is to maximize their revenue which is generated by the sale of (or providing access to) their information to the interested buyers. The objective of the buyers, apart from purchasing this information or purchasing the access to this information, is to be able to obtain the "in demand" information bundles from the maximum users in the most profitable transaction.

We acknowledge that not all users may possess the same information bundle and also that not all buyers want the same types of information bundle.

The first step then, is to understand the demand for and the value of this information bundle currently in the information market. One of the most suitable techniques to gauge or estimate the demand is to allow users to quote their (maximum)

price for a particular information bundle. This process would give us a conservative estimate of the demand in the information market for a particular information bundle.

The factors that would determine the demand for the information bundle in the information market are:

- amount that each interested buyer quotes for an information bundle
- the number of interested buyers for an information bundle
- depending on the scenario, how fast does this demand change, i.e. time

In a markov decision process, when a state change occurs, it rewards the 'decision maker' with a certain reward for choosing that state change.

In our scenario, the policy to be chosen by the user (to sell his information or not) would be determined by the reward for the user which in this case is the revenue generated by the potential sale of this information. This reward in turn is determined by the price being paid by the potential buyer (that signifies a buyer's individual interest), the amount of information a buyer is looking for and the total number of buyers quoting their prices for a particular information bundle (that signifies the demand in the market). (Refer Table 7.1 for the parameter list).

Let the number of users in the market be $\boldsymbol{j}$. Let us also assume that the 'information bundles' in possession by the users are $\boldsymbol{n}$. Now the demand for each information bundle is $\boldsymbol{\lambda_n}$ which is calculated based on the buyer's initial quote $\boldsymbol{b_{in}}$ and the number of buyers who are interested in that information bundle. Each of these buyers would need a certain 'amount' of information $\boldsymbol{x_{in}}$, either one day or one month (for the sake of convenience, we state this number of days). Let each user's private valuation of his information bundles be $\boldsymbol{p_n}$.

Our goal then is to maximize each user 'j's' revenue which is 'R'.

$\boldsymbol{R_j = \Sigma c_i d_i}$

where, $\boldsymbol{c_i = \Sigma_{n \in N} b_{in} x_{in} \lambda_n}$

Now if each user $\boldsymbol{j}$ is satisfied with the revenue $\boldsymbol{R_j}$ for his information bundles, then there is a higher likelihood that he will sell or give access to his information,

$\boldsymbol{\Sigma_{n \in N} p_j \geq R_j}$

i.e. if the revenue generated is more than the expected private valuation then the user would choose to sell his information.

$\boldsymbol{\Sigma_{n \in N} p_j < R_j}$

And if the revenue does not satisfy the private valuation of the user, then he may choose to not sell or share his information in the information market.

Table 7.1: Parameter list

Parameter	Stands for
c_i	cost to every buyer i
d_i	1 (if buyer i is purchasing that information) 0 (otherwise)
n	The information bundle of a user
b_{in}	initial price from buyer i
x_{in}	Amount of information (eg: one days worth of information) the buyer is interested in
λ_n	Demand for information bundle per unit time
p_n	Private valuation of the user for each information bundle

7.3 Results & Discussion

We ran simulations with 5 Buyers and 50 users and 5 information bundles (A,B,C,D,E), with each user having different information bundles and each buyer bidding for different bundles. The prices for each information bundle varied according to the demand for that information bundle. And accordingly, the revenue for each user with that information bundle varied.

The individual prices quoted by the buyers along with the amount of the bids from each of the buyers indicates their willingness to pay, i.e. their desire to purchase a particular information bundle. This is indicative of the demand for the information bundle. Figure 7.1 shows us the demand for each information bundle in our simulation. We have 5 information bundles and based on the number and amount of the bids for each of the information bundles, we can see that the demand for information bundle B is the least and that for information bundle E is the maximum.

Not all the buyers request for the same information bundle. Figure 7.2 displays

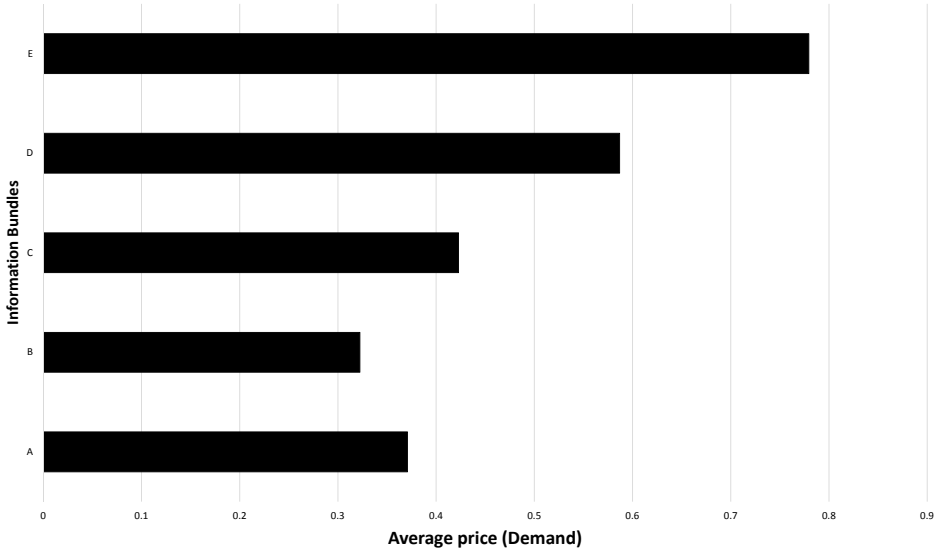


Figure 7.1: Demand for the information bundle

the information bundles that different buyers would need and also the corresponding cost to each of the buyers for their requested bundle (per user). From the bar graphs we see that though the demand for information bundle 'A' and 'E' is made by all 5 buyers, the willingness to pay for the respective bundles by each of them varies. This in turn drives the high demand for information bundle 'E' (seen in Figure 7.1).

The line graph in the Figure 2 shows the cost to the buyers based on the demand for the information bundles and the individual prices quoted by the buyers.

Figure 7.3 shows us the revenue that is generated for each user who possesses a different suite of information bundles. Each user has then the potential to generate a certain revenue by the sharing or the sale of his or her information. The differing levels of revenue generated correspond to the real world scenario where different buyers have different prices for goods depending on their willingness to buy that good. Here, depending on the buyer's willingness to buy an information bundle, the price paid also similarly differs. But it follows the demand trend in the market (shown in Figure 7.1). As seen in Figure 7.3, the users with no information bundles (NULL) do not have any revenue.

The revenue generated, based on the proposed model, is generated by the demand for an information bundle. Though this is an often used economic concept, it's

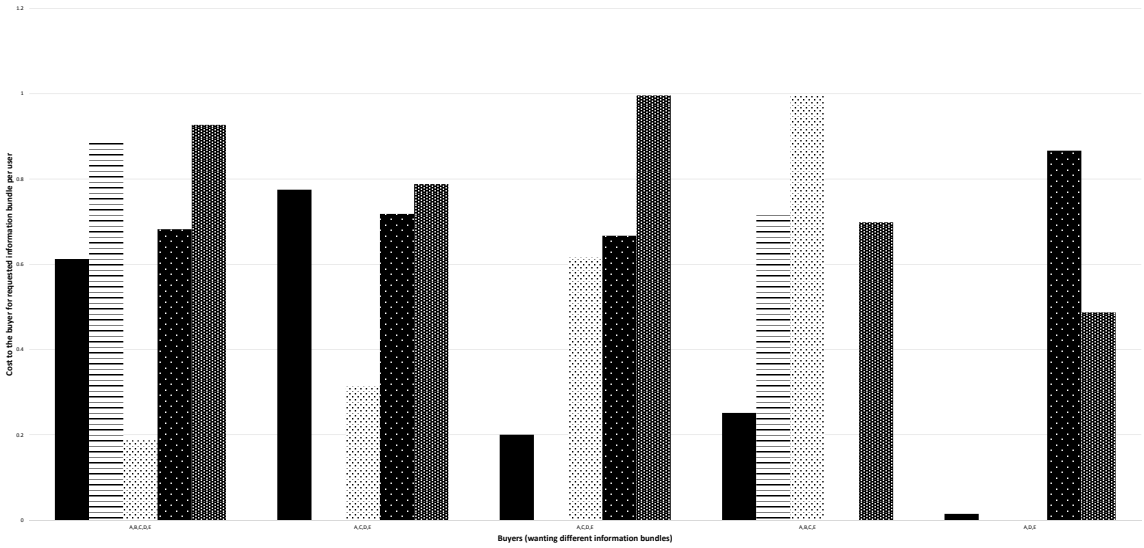


Figure 7.2: Cost to the buyer per user for the requested information bundle

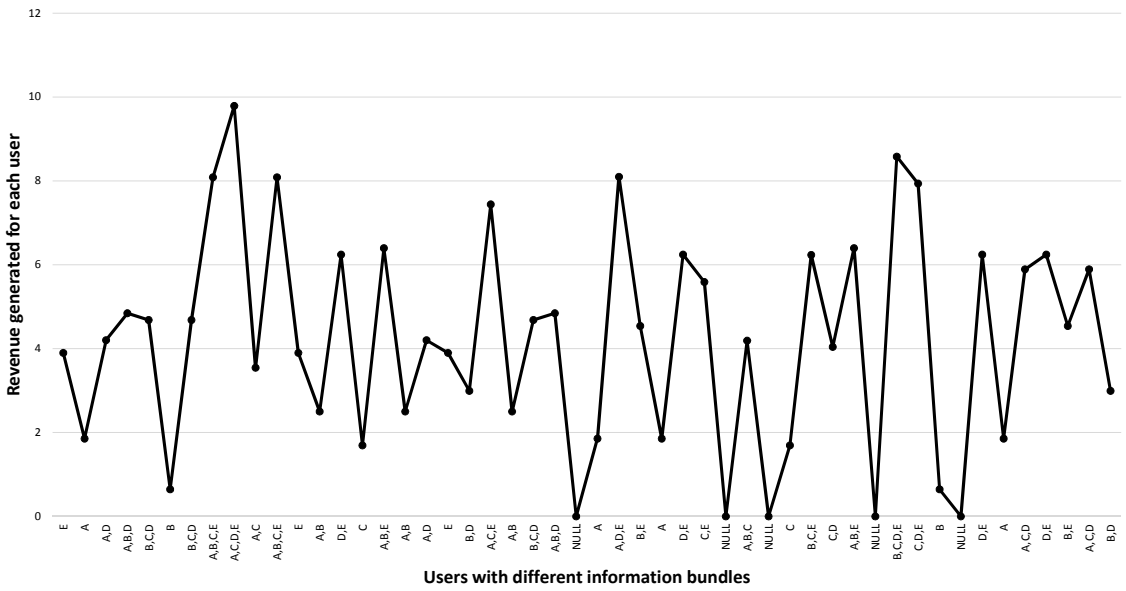


Figure 7.3: Revenue generated for the users

implementation in the information market presents a transparent method for buyers to purchase information from users directly without the involvement of any third parties, as is happening currently in the form of data brokers. Thus, the proposed model provides the foundation with a simplistic solution to the problem of pricing

information which can then be expanded further. These results demonstrate that the implementation of an application in the form of a mobile app or a browser plug-in is a realistic and a practical solution to the question of generating revenue for the user information.

7.4 Summary

With the increased proliferation of social media and the ever rising number of internet users who are constantly sharing information, (about themselves or otherwise) through the platform of social media, organizations have realized that the way to generate more profit is to capitalize on this resource. This has lead to alarming problems of privacy infringement and surreptitious trading of information without the internet user's knowledge. This has given rise to problems of price discrimination and profiling.

One of the most realistic and practical ways to make users aware of the privacy infringement by data brokers is to inform them of the value of their information. For this, having appropriate pricing mechanisms for the same is the way ahead. These pricing mechanisms must work to empower the internet users by giving them more control about the information they can sell and share and information they want to keep private. For this reason there is a need for platforms as described in this paper which will then go on to eliminate the need for data brokers who sneak around and collect user information.

In this chapter, we presented a pricing model by applying the Markov process as a means to arriving at the pricing problem. Using the Markov process as a base, we were able to create our revenue generation model for the user using which the user can avail himself of the monetary benefits that come with selling his information to prospective buyers. We have attempted to provide a balanced model by incorporating the buyer's point of view, specifically the price has been adjusted keeping the current demand in the market for an information bundle. The combination of the markov process along with the economic concept of demand has resulted in the creation of a fair and simplistic pricing model.

This pricing model was implemented and tested for its viability through the means

of computer simulations. It was seen that the model provides a realistic pricing scheme for the scenario of selling information in the information market by a user. It connects the seller, which in this case is our internet user, and the buyer directly and eliminates the need for any kind of "middle man" type of entity who may be manipulated.

8

Pricing based on Portfolio Optimization

Today, with the combined power of the cloud, big data has come to include data from all sources, right from social networks to a person's shopping history. For organizations, this is a wealth of information that can be analyzed and exploited for their own benefit and this is generally carried out without a thought about the privacy violations or compensating the users. In this chapter, we have attempted to look at this concept of an "information market" and have ventured to ascertain the value of the big data by involving users and incentivize them to share their information by rewarding them with monetary compensation for their information. We have attempted to perform this task by treating user information as a commodity or an asset which can be traded in the information market utilizing the concept of portfolio optimization which can help buyers decide on the types of information they want to "invest" in, which would then be dispersed to the users.

In our model, the information price reflects the demand in the market for a particular type of information. This in turn allows the suppliers of information (the ordinary internet users) to realize the value of their information and grants them control over the type of information they would like to share and monetize. Our problem then, is to find the optimal portfolio based on various factors like the demand for the information, the user's sensitivity with parting with that information and so on.

8.1 Information pricing model

Information shares the classic characteristics of a typical asset. It has the ability to achieve a positive economic flow, the access to information can be controlled by the owner of the information and the ownership of information is an event that has already occurred. Thus, information being similar in characteristic and importance like an asset, we have borrowed from the idea of portfolio optimization, specifically the Markowitz mean - variance model to devise a suitable portfolio for the buyer to invest. Markowitz formulated the fundamental theorem of a mean variance portfolio framework [38], which explains the trade-off between the expected returns(mean) and risk of a portfolio(variance). The Markowitz model is not just restricted to the financial markets and has been applied in diverse scenarios ranging from electricity market [53] to increasing forest resource [85].

We have adapted the mean - variance model to cater to the information market to develop the information pricing model. The primary goal of the *information pricing model* is to provide a platform for the buyers to choose the information they would want to purchase. Through the medium of this platform, the buyer then proceeds to 'invest' in a particular set of assets, *i.e.* the information type of his choosing referred to as 'information bundle' (sets of personal and/or private information of the user and could be supplemented with insights based on that personal information). This would form the portfolio for each buyer. After the choosing of the optimum portfolio, the buyer can then determine the number of users from whom he wants to purchase this information from. In our model we have a set of users who wish to sell

their information and at the other end we have a set of buyers who are interested in purchasing this information. The objective of the users is to maximize their revenue which is generated by the sale of this information to the interested buyers.

8.1.1 Proportion of Investment

The first step is to get the proportion invested in an asset by the buyer. The buyer specifies the proportion of investment (in weighted percentage) that he would like to invest in each information bundle. Let m_{ij} represent the input of each user 'j' for the information bundle 'i'. This step demonstrates the demand in the market for the information bundle. Analyzing this demand allows us to understand the correlation between the demand and the corresponding supply in the market for the information bundle (by the users) in the market.

8.1.2 Incorporating the risk

The covariance allows us to quantify the risk. This risk is two fold - it represents the hesitancy the user feels while parting with his more sensitive and private information and it also represents the buyer's risk and uncertainty of whether he will be able to purchase and make use of the information. The more sensitive and private and possibly invasive the information bundle in question, the more is the risk. We assign this risk in the range [-1,+1] with -1 being the most risk involved and 1 being the least risk. If the information is of a personal and sensitive type, then the risk involved in the potential sale of this information is high. If the information is generic and demographic the risk is definitely lower. Hence we have $r_i \propto$ (information sensitivity) where r_i is the risk with 'i' representing the information bundle.

8.1.3 Rate of return - for the buyer

Generally, the rate of return of an asset is calculated by determining the change in the value invested. But for information, the rate of return of information cannot be determined in such a straight forward manner. To determine the rate of return of information from the point of view of the buyer, the buyer needs to look into it from the point of view of 'value of information'. The optimal method to determine the rate of

return of information would be the one that would give the highest 'Expected Monetary Value'. This would mean that the buyer would need to choose the information bundle that would allow him to make the most profit using the information bundle. Let this be $\boldsymbol{v_{ij}}$ for every buyer 'j'.

8.1.4 The Information Pricing model

In this case each buyer 'j' can use the following method based on the mean variance model to 'invest' in the assets of their choice to build up a profitable portfolio:

$$\mathbf{min} \sum_{i=1}^{n} \sum_{k=1}^{n} \boldsymbol{m_{ij} m_{kj} r_{ik}}$$

such that:

$$\sum_{i=1}^{n} \boldsymbol{m_{ij} v_{ij}} = \boldsymbol{R^*}$$

Here $\boldsymbol{R^*}$ is the expected level of returns each buyer 'j' hopes to achieve with his investment.

$$\sum_{i=1}^{n} \boldsymbol{m_{ij}} = \mathbf{1}$$

$$\mathbf{0} \leq \boldsymbol{m_{ij}} \leq \mathbf{1} \qquad \boldsymbol{i} = \mathbf{1, 2, ...} \boldsymbol{n}$$

The above constraint is to make sure that all the investments are utilized. Our model can help the buyer ascertain his most profitable portfolio on information bundles to invest in. Once the buyer can ascertain his most profitable portfolio, he can specify the number of information bundles he requires, i.e. the number of users from whom this information is required from. The total amount invested by each buyer would then be appropriately distributed amongst the users, i.e. the sellers of the information.

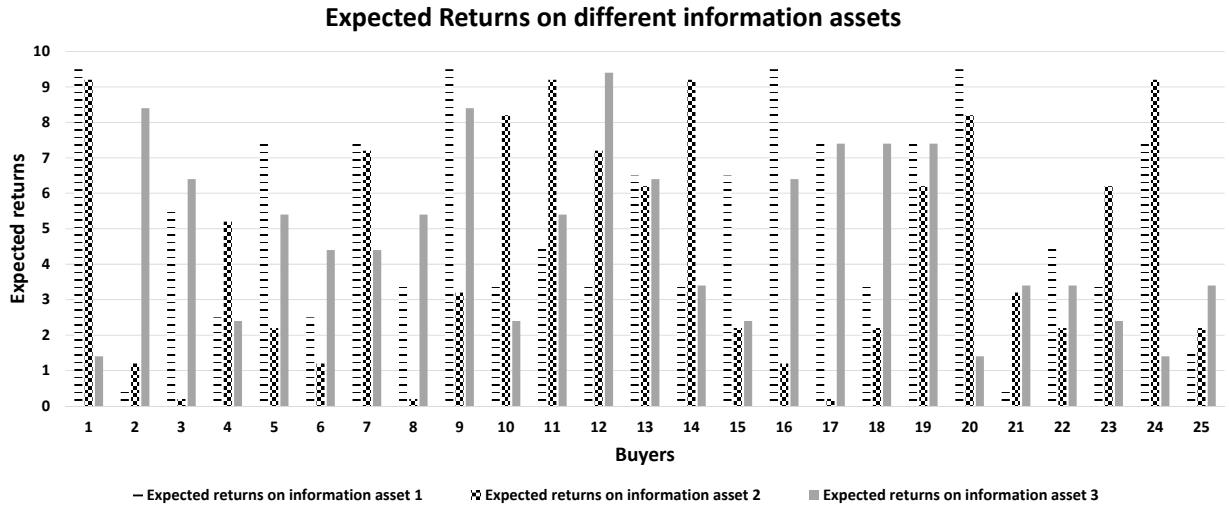


Figure 8.1: Expected returns of the buyers

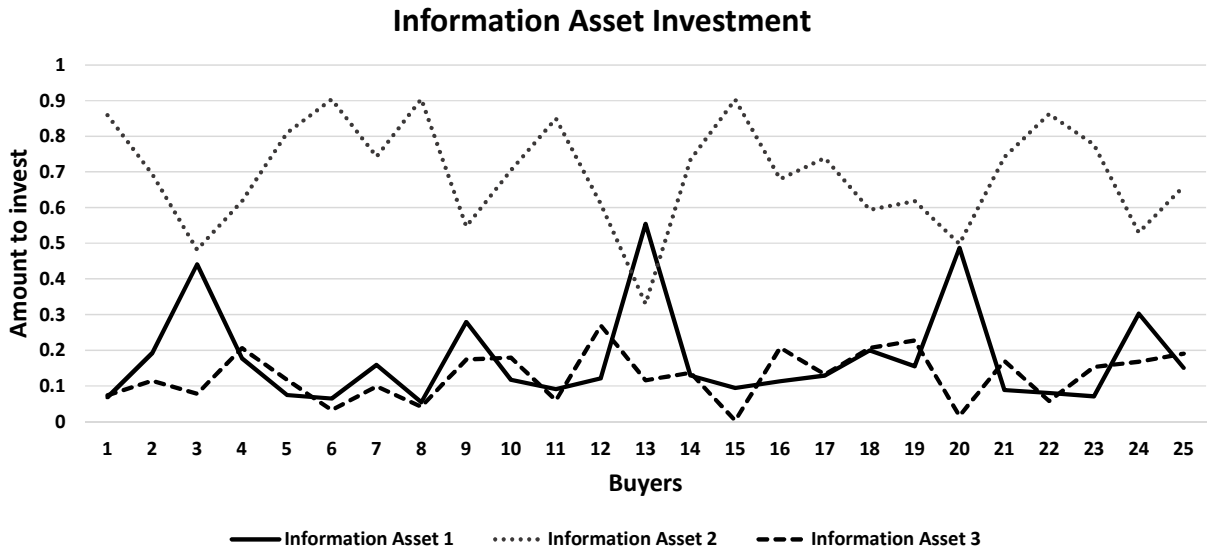


Figure 8.2: Information Investment for the buyers

8.2 Results & Discussion

We solicited the help of 25 participants to take on the role of buyers of information. We explained our scenario to them and showed them the various information bundles available. Now each of these buyers would be expecting different rates of return from their investment. These buyers look at the risk associated with each information category and decide their investment amounts for the information categories. The only stipulation is that the investments should be add up to 1 (keeping in mind the working of the model). Of the assets presented to the buyers, information asset 2 can be considered as a high risk asset because of its sensitive nature (eg: Income, Health information etc). Figure 8.1 shows us the expected rates of return that the buyers would expect from their investments. In the graph we can see that buyer 2 expects a high return from asset 3 and a relatively low return on asset 2. From the Figure 8.2, which is the results available after running the inputs from the buyers, we see that buyer 2 to obtain his expected return on asset 2, would need to invest a higher amount as opposed to what he would need to invest for asset 3. This shows us that the asset 2 is a far in demand and as compared to assets 1 and 3 and because it is a "high risk" asset, it requires a relatively higher initial investment. The Information Pricing model described in this paper is driven by the economic principal of demand in the market - the more the demand, the better the price. There are no hidden costs or any other extraneous costs added for the buyer. This also gives the sellers of information (the internet users)more control over what personal and private information they would like to sell. This is unlike what is currently followed by data brokers who have their own techniques for pricing along with extra costs.

8.3 Summary

In this chapter, we have treated user information as an asset in the Internet information market and have thus developed the 'Information pricing model' inspired by the technique of portfolio optimization. Using this model, the buyer can objectively assess which information type the buyer would want to "invest" in. This gives the buyer

incentive to participate in the process as the model gives the buyer better management over which information that he wants to purchase. This is not available to the buyer in the current scenario with the presence of data brokers. Based on our preliminary results we believe that this provides a fair and balanced mechanism to tackle the question of information pricing in a simplistic and practical manner.

9

Conclusions & Future Work

Most organizations today have realized the potential of the revenue generating capacity of user information. Though users do sign away any rights to their information when they check the "terms and conditions" box, because of the lengthy legalese in which this is framed, the users do not fully grasp the importance of what they are giving away.

With the increased proliferation of social media and the ever rising number of internet users who are constantly sharing information, (about themselves or otherwise) through the platform of social media, organizations have realized that the way to generate more profit is to capitalize on this resource. This has lead to alarming problems of privacy infringement and surreptitious trading of information without the internet user's knowledge. This has given rise to problems of price discrimination and profiling.

With the rising market in the wearable devices, companies like Google also have started investing in technology that utilizes user information for the purposes of ad and product placement (Pay-Per-Gaze patent). We also have instances of companies using the data from their employees fitness devices. Apps like "Tinder" are also performing analysis on available user data surreptiously [1]. There is a huge market that exploits user information on social media for the pure purpose of inundating the users with ads tailored to their liking based on the user information on the social media which is surreptiously used by data companies. According to a Federal Trade Commission (FTC) report [37], data brokers, generate around 200 billion dollars in revenue annually. This business of making money off of user information should be a source of concern for the users.

Information is being shared at an unprecedented rate today. And for the organizations, all this information and the insights from this information, which make up it's big data, are the biggest asset. Any and all information about you is the meal ticket for these data brokers to earn a huge amount of money. And the users just seem apathetic to this obvious infringement of their information and their privacy.

The value of an object is more easily understood to a layman when it is put in terms of a monetary value. One of the most realistic and practical ways to make users aware of the privacy infringement by data brokers is to inform them of the value of their information. For this, having appropriate pricing mechanisms for the same is the way ahead. These pricing mechanisms must work to empower the internet users by giving them more control about the information they can sell and share and information they want to keep private. For this reason there is a need for platforms as described in this paper which will then go on to eliminate the need for data brokers who sneak around and collect user information.

9.1 Conclusions

In this thesis, we presented pricing models to tackle the question of how to price user information. The discussed pricing models were developed keeping in mind various

[1] http://www.theverge.com/2016/1/14/10773026/stolen-twitter-game-pulled-from-app-store

scenarios and concepts with the common thread of balancing privacy of the user along with the fair valuation of the information and the estimation of the pricing of the information for the buyer.

We started this thesis with our preliminary findings about the concept of information pricing. We added to these findings with any relevant research in the related fields of privacy and the existence of the current pricing models. This provided us with a robust basis for our research thesis.

In Chapter 3, we presented a brief overview of our work accompanied by a short case study that explored the need for the presence of pricing models. This case study helped us isolate our research goals and helped us focus on the various scenarios for our pricing models.

We then proceeded to deal with the privacy incorporated pricing where we balance a user's privacy in the pricing model. This scenario was dealt with in different techniques in Chapters 4 and 5.

In chapter 4, we have provided for the user‘s concerns about privacy and at the same time provided for a means for the user to sell this information. This method provides a fair and balanced mechanism to tackle the question of information pricing in a simplistic and practical manner.

The scenario in this chapter targets numerical information. The problem of protecting privacy while sharing other types of information categories is a possible avenue for further research. Our method allows users to judge for themselves the demand for different types of information and decide for themselves which information they want to reveal. Our pricing model allows for user privacy by obfuscating the user information and adjusts the pricing accordingly. Our model does not in any way depend on the number of the users or buyers and thus can be scaled as desired. Using our model eliminates the need to use any third party and connects the buyers and the users (the sellers of information) directly to each other.

In chapter 5, our method is directly related to the information each user wants to sell in the online information market place. In this chapter, we have treated user

information and the big data generated from it as an asset in the Internet information market. To assuage the user of the privacy of his information, we have incorporated the obfuscation of the user information by the introduction of noise taken from a normal distribution. Now for the buyer, this obfuscated information is a risk for the buyer because the buyer is getting obfuscated information. This risk calculation is done by applying the concept of the Sharpe ratio. The Sharpe ratio is used to give a risk adjusted measure that can help the buyer calculate the return that they would get when they purchase this obfuscated information.The Sharpe ratio thus provides a fair measure for the buyer to assess the quality of information and decide on the prices for the obfuscated information categories.

The above described method caters to the concerns of the stakeholders involved in a fair and balanced fashion. The usage of the Sharpe ratio as a risk adjusted measure allows the buyer to know the average return he can expect from the information he wishes to purchase while at the same time the user being assured of a certain amount of privacy can sell his information to the prospective buyers without any hesitancy or misgivings.

We then provided a scenario for a less privacy conscious individual user and for those buyers who would not want to purchase obfuscated information. We dealt with this scenario by developing pricing models by using diverse concepts. In chapter 6 we have presented our pricing model that allows a user to realize the value of his information and the potential revenue he can make off of it. In the scenario described above, we presented the information using the concept of Shannon's information theory to understand the value of information that is lost when it is revealed. We also presented the demand analysis using two methods - the traditional method and the exponential mechanism method. In the traditional method, there is a possibility of untruthfulness and collusion that may cause the pricing of information to not be calculated in the optimal way. To counter this, we have tested the the demand by applying the exponential mechanism which ensures that the buyers are truthful and honest in stating their demand. The exponential demand method provides the buyers with the least incentive to lie about their true valuations and thus enables a

more accurate pricing for the information in question. The reflection of adopting the exponential mechanism on the user revenue is minuscule and gives the user more trust to participate in our pricing model.

And in chapter 7, we presented a pricing model by applying the Markov process as a means to arriving at the pricing problem. Using the Markov process as a base, we were able to create our revenue generation model for the user using which the user can avail himself of the monetary benefits that come with selling his information to prospective buyers. We have attempted to provide a balanced model by incorporating the buyer's point of view, specifically the price has been adjusted keeping the current demand in the market for an information bundle. The combination of the markov process along with the economic concept of demand has resulted in the creation of a fair and simplistic pricing model.

In chapter 8 describes the information pricing model from the point of view of the buyer by allowing to help him choose the information to invest in and how much to invest in

This pricing model was implemented and tested for its viability through the means of computer simulations. It was seen that the model provides a realistic pricing scheme for the scenario of selling information in the information market by a user. It connects the seller, which in this case is our internet user, and the buyer directly and eliminates the need for any kind of third party conduit who stands the risk of being manipulated.

When we started out with our research, we had certain research goals in mind as described in Chapter 1. By the end of this thesis, we can say that we have achieved the same as elaborated below:

1. We have developed pricing mechanisms for the scenarios of:

    - Privacy aware user and tolerant buyer - The user is an individual who values his privacy but would still like to sell his information and the buyer is tolerant enough to accept the user information with certain obfuscations. The idea being that the buyer would compensate the user based on the quality of the obfuscated information.

- Privacy insensitive user and an intolerant buyer - The user is not a very privacy conscious individual and the buyer is willing to accept only unadulterated and unobfuscated information.

2. Our pricing mechanisms achieve the fair balance between the needs of all the stakeholders by considering their requirements and adjusting the pricing accordingly.
3. Our models have the advantage of not only being practical but also easily integrate-able with any business that has an online presence. A model that fits into an existing platform and does not affect the functioning of the platform is more readily adopted than a stand alone model.

9.2 Comparison with existing work

Currently, the pricing models that exist for the data market can be divided based on the type and ownership of the data. One can be referred to as "data markets" and other as "personal data markets". Data markets are marketplaces where the ownership of the data is with the entities actually collecting the data, for eg,: the data that a company collects as part of its registration process. Though this data is about the users, it isn't owned by the users and thus never see the profits from its subsequent monetization. Examples of these types of commercial data markets are plenty, one of them being Qlik [9] and the soon to be retired Windows Azure Marketplace [2]. So in this scenario, the entire data of a whole set of users is located in one place (so to speak) as a database and can be repeatedly queried. The person querying (referred to as the buyer) has to pay either a fixed amount for a year (a flat fee) or pays along for each query. There are solutions that address the pricing in these situations by suggesting different pricing models that can be cost effective for the buyer.

This is different from our problem because we do not collect user information and store it in a centralized database. In the above case, the users to whom this data belongs to and is about do not get an opportunity to monetize on their information.

[2]https://rcpmag.com/articles/2016/11/28/microsoft-shutters-azure-datamarket.aspx

Also, many times the buyers hesitate with their queries since the queries could reveal information that the buyer would rather keep private.

Personal data markets on the other hand, though still in their nascent stage, are slowly coming up as a means for users to monetize on their information. The most relevant examples are websites like Meeco [8] which require users to sign in and give access to their information and in return they act as vaults and every time an application requires access to a users information, the user gets paid. Or the example of 'Datacoup' [7] with their scheme of paying users for access to their personal data using social media and their credit or debit card feeds. Though they strive to make this information devoid of any personally identifiable information, the control of this information and any subsequent profits from this information is wrested away from the users. The issue with such a solution is that the users have to still allow such applications access to their information. This at a later stage could be misused and the users wouldn't have much idea about this.

The other existing technique is the application of differential privacy [50]. But this in turn puts the buyers at a backfoot since they will end up having to pay more money for any usable information. Yet another comparison could be made with the idea of applying game theory to allow users to obtain a certain pay off based on their decision [71].But this means that the users only gain if their decision matches the "winning" decision. And in all other cases, they stand to lose. Another point is that this theory to the best of our knowledge has been applied to only specific scenarios. Another technique to price use information is auctions. As explored by [111] the auctions are based on the exponential mechanism. But this puts the onus of setting the price solely on the buyers which could tend to put the users at a loss. And finally the "middle man" approach [52] where in a third party is present between the user (seller) and buyer who collects the information from the users and then decides the price. Again in this case the control is wrested away from the user.

In our scenario we would like the users to retain control over the type of information they would like to share and be able to monetize on this information without having

to lose control over this information. This is the reason we have also discussed a privacy protecting model where the user need not share his actual raw data but just a statistic. While we realize that all the models do not apply to all scenarios of interested buyers, we feel that some models could work for certain situations and some for other situations.

Our goal in developing these models has been to be able to create not just awareness of the monetization of information but also to make the user realize that he can have control and privacy over his information while still being able to capitalize on it without having to depend on a third party to intervene and set prices and gain access to their information.

In our work, we have presented another perspective of looking at the problem of information pricing and have tried to balance the needs of all the stakeholders involved. In comparison with the others,

- There is no involvement of any "middle man" who would collect user information
- The pricing is not skewed towards one side, rather it is fair and balanced
- Users retain control over their information while also being aware of the value of their information

As elaborated in our future work, we hope to take this momentum forward and look about the actual implementation of these models for different scenarios so that users can avail of the opportunity of monetization over their information.

9.3 Future Work

This thesis covers the issue of information pricing and presents pricing models that provide for a balanced information pricing suiting to the needs of all the stakeholders in the transaction. We feel that there are certain aspects that can be further researched and integrated into the above presented pricing models.

- Integrity - For the purpose of this thesis, we have worked under the research assumption that the integrity of the information is true. This means that we have assumed that the user sharing the information is actually the owner and

the information is about the user. But there are instances like 'Stolen', an app that allowed anyone to sell anyone's twitter information and feed (due to increased complaints, this app was pulled from the app store [3]) and websites like 'Slur' [4] that encourages the selling of information not necessarily belonging to the person selling the information. In light of these, we feel that there should be some measure to integrate the idea of integrity into the pricing model itself. The research on integrity looks at ideas like watermarking or work it into the XML data. This could be one of the ideas to expand our pricing models.

- Historical/time sensitive pricing - For traditional goods, the passage of time has a significant effect on the prices. When information is looked as a good, time could have an effect on the value and pricing. Certain types of information could be more valuable over time (like a person's credit history) but certain types of information may depreciate in its value over time (like a person's educational institutions).

- Integration with other real world applications & services - In the real world, information of various types would be encountered - images, video, string, character etc. For the purpose of this thesis, we have handled only information of the numeric nature. The application of the techniques discussed in this thesis may not be suitable for the estimation of the value of information belonging to the various other datatypes. Handling multiple types of data encountered in the real world would allow for the incorporation of pricing models within real world applications. The presence of such models that assure the user of monetary compensation could be of immense value to all types of businesses having an online presence. As shown in our case study in Chapter 3, users are also willing to accept this monetary compensation in terms of store credit and discount vouchers. This has tremendous implications for businesses and organizations who are looking to not just retain their existing customers but also are looking to attract new customers. If assured of monetary compensation, the customers

[3] http://www.theverge.com/2016/1/14/10773026/stolen-twitter-game-pulled-from-app-store
[4] http://slur.io/

would in turn keep returning to the businesses for their service.

Bibliography

[1] http://removeandreplace.com/2013/03/13/how-much-data-is-on-the-internet-and-generated-online-every-minute/.

[2] http://media.kaspersky.com/pdf/it-risks-survey-report-cost-of-security-breaches.pdf.

[3] https://www.preceden.com/timelines/47418-evolution-of-cloud-storage.

[4] http://www.cioreview.com/cxoinsight/cloud-storage-evolution-of-real-estate-on-the-information-superhighway-nid-10695-cid-117.html.

[5] http://www.cnet.com/how-to/onedrive-dropbox-google-drive-and-box-which-cloud-storage-service-is-right-for-you/.

[6] http://www.technologyreview.com/news/524621/sell-your-personal-data-for-8-a-month/.

[7] Datacoup. https://datacoup.com/.

[8] Meeco. https://meeco.me/.

[9] Qlik. http://www.ometis.co.uk/products/datamarket.

[10] A. Acquisti. Privacy in electronic commerce and the economics of immediate gratification. In *Proceedings of the 5th ACM conference on Electronic commerce*, pages 21–29. ACM, 2004.

[11] A. Acquisti. Nudging privacy: The behavioral economics of personal information. *Digital Enlightenment Yearbook 2012*, page 193, 2012.

[12] A. Acquisti, R. Dingledine, and P. Syverson. On the economics of anonymity. In *Financial Cryptography*, pages 84–102. Springer, 2003.

[13] A. Acquisti, A. Friedman, and R. Telang. Is there a cost to privacy breaches? an event study. In *WEIS*, 2006.

[14] A. Acquisti and R. Gross. Imagined communities: Awareness, information sharing, and privacy on the facebook. In *Privacy enhancing technologies*, pages 36–58. Springer, 2006.

[15] A. Acquisti and J. Grossklags. Privacy and rationality in individual decision making. *IEEE Security & Privacy*, 2:24–30, 2005.

[16] B. Aiello, Y. Ishai, and O. Reingold. Priced oblivious transfer: How to sell digital goods. In *Advances in CryptologyEUROCRYPT 2001*, pages 119–135. Springer, 2001.

[17] R. Anderson and T. Moore. The economics of information security. *Science*, 314(5799):610–613, 2006.

[18] Aricent and F. Design. Smart home opportunity: Balancing customer data and privacy, 3, February 2010.

[19] F. Baccelli and J. Bolot. Modeling the economic value of the location data of mobile users. In *INFOCOM, 2011 Proceedings IEEE*, pages 1467–1475. IEEE, 2011.

[20] S. Bandyopadhyay. Antecedents and consequences of consumers online privacy concerns. *Journal of Business & Economics Research (JBER)*, 7(3), 2011.

[21] B. J. Bates. Information as an economic good: A re-evaluation of theoretical approaches. *Mediation, information, and communication. Information and behavior*, 3:379–394, 1990.

[22] R. Bellman. A markovian decision process. Technical report, DTIC Document, 1957.

[23] A. R. Beresford, D. Kübler, and S. Preibusch. Unwillingness to pay for privacy: A field experiment. *Economics Letters*, 117(1):25–27, 2012.

[24] S. Berthold and R. Böhme. Valuating privacy with option pricing theory. In *Economics of information security and privacy*, pages 187–209. Springer, 2010.

[25] S. Binitha and S. S. Sathya. A survey of bio inspired optimization algorithms. *International Journal of Soft Computing and Engineering*, 2(2):137–151, 2012.

[26] R. Böhme and D.-I. S. Koble. Pricing strategies in electronic marketplaces with privacy-enhancing technologies. *Wirtschaftsinformatik*, 49(1):16–25, 2007.

[27] E. Bonabeau and C. Meyer. Swarm intelligence: A whole new way to think about business. *Harvard Business Review*, 79(5):106–115, 2001.

[28] J. Bonneau and S. Preibusch. *The privacy jungle: On the market for data protection in social networks*, pages 121–167. Springer, 2010.

[29] P. Booth, P. Gaskell, and C. Hughes. The economics of data: quality, value & exchange in web observatories. In *Proceedings of the 22nd international conference on World Wide Web companion*, pages 1309–1316. International World Wide Web Conferences Steering Committee, 2013.

[30] J. L. Boyles, A. Smith, and M. Madden. Privacy and data management on mobile devices. *Pew Internet & American Life Project*, 4, 2012.

[31] S. Buffett, M. Fleming, M. Richter, N. Scott, and B. Spencer. Determining internet users' values for private information. 2004.

[32] O. Candogan, K. Bimpikis, and A. Ozdaglar. Optimal pricing in social networks. *ACM SIGecom Exchanges*, 10(3):15–17, 2011.

[33] A. Caracas and J. Altmann. A pricing information service for grid computing. In *Proceedings of the 5th international workshop on Middleware for grid computing: held at the ACM/IFIP/USENIX 8th International Middleware Conference*, page 4. ACM, 2007.

[34] J. P. Carrascal, C. Riederer, V. Erramilli, M. Cherubini, and R. de Oliveira. Your browsing behavior for a big mac: Economics of personal information online. In *Proceedings of the 22nd international conference on World Wide Web*, pages 189–200. International World Wide Web Conferences Steering Committee, 2013.

[35] Y. Chen. Information valuation for information lifecycle management. In *Autonomic Computing, 2005. ICAC 2005. Proceedings. Second International Conference on*, pages 135–146. IEEE, 2005.

[36] E. Coiera. Information economics and the internet. *Journal of the American Medical Informatics Association*, 7(3):215–221, 2000.

[37] F. T. Commission et al. Data brokers: A call for transparency and accountability. *May. www. ftc. gov/system/files/documents/reports/data-brokers-call-transparency-accountabilityreport-federal-trade-commission-may-2014/140527databrokerreport. pdf*, 2014.

[38] T. Cura. Particle swarm optimization approach to portfolio optimization. *Nonlinear Analysis: Real World Applications*, 10(4):2396–2406, 2009.

[39] C. DESMARAIS. Who is gathering your personal information?, 11 June 2015. retrieved June 20, 2015.

[40] H. Du and M. N. Huhns. A multiagent system approach to grocery shopping. In *Advances on Practical Applications of Agents and Multiagent Systems*, pages 195–200. Springer, 2011.

[41] E. Dumbill. Making sense of big data. *Big Data*, 1(1):1–2, 2013.

[42] C. Dwork. Differential privacy. In *Automata, languages and programming*, pages 1–12. Springer, 2006.

[43] V. Erramilli. The tussle around online privacy. *IEEE Internet Computing*, 16(4), 2012.

[44] I. V. Evstigneev, T. Hens, and K. R. Schenk-Hoppé. Mean-variance portfolio analysis: The markowitz model. In *Mathematical Financial Economics*, pages 11–18. Springer, 2015.

[45] M. B. E. Fayek, I. A. Talkhan, and K. S. El-Masry. Gama (genetic algorithm driven multi-agents) for e-commerce integrative negotiation. In *Proceedings of the 11th Annual conference on Genetic and evolutionary computation*, pages 1845–1846. ACM, 2009.

[46] E. J. Friedman and D. C. Parkes. Pricing wifi at starbucks: issues in online mechanism design. In *Proceedings of the 4th ACM conference on Electronic commerce*, pages 240–241. ACM, 2003.

[47] Y. Gao, G. Zhang, J. Lu, and H.-M. Wee. Particle swarm optimization for bi-level pricing problems in supply chains. *Journal of Global Optimization*, 51(2):245–254, 2011.

[48] S. Garnier, J. Gautrais, and G. Theraulaz. The biological principles of swarm intelligence. *Swarm Intelligence*, 1(1):3–31, 2007.

[49] A. Gershkov and B. Moldovanu. Dynamic allocation and pricing: A mechanism design approach. *International Journal of Industrial Organization*, 30(3):283–286, 2012.

[50] A. Ghosh and A. Roth. Selling privacy at auction. *Games and Economic Behavior*, 2013.

[51] P. Gill, V. Erramilli, A. Chaintreau, B. Krishnamurthy, K. Papagiannaki, and P. Rodriguez. Best paper–follow the money: understanding economics of online aggregation and advertising. In *Proceedings of the 2013 conference on Internet measurement conference*, pages 141–148. ACM, 2013.

[52] V. Gkatzelis, C. Aperjis, and B. A. Huberman. Pricing private data. *Available at SSRN*, 2012.

[53] F. Gökgöz and M. E. Atmaca. Financial optimization in the turkish electricity market: Markowitz's mean-variance approach. *Renewable and Sustainable Energy Reviews*, 16(1):357–368, 2012.

[54] R. Gross and A. Acquisti. Information revelation and privacy in online social networks. In *Proceedings of the 2005 ACM workshop on Privacy in the electronic society*, pages 71–80. ACM, 2005.

[55] K. P. Gummadi, B. Krishnamurthy, and A. Mislove. Addressing the privacy management crisis in online social networks. In *WWW (Companion Volume)*, pages 841–842, 2013.

[56] A. Gupta, K. Ligett, F. McSherry, A. Roth, and K. Talwar. Differentially private combinatorial optimization. In *Proceedings of the Twenty-First Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1106–1125. Society for Industrial and Applied Mathematics, 2010.

[57] S. Ha, S. Sen, C. Joe-Wong, Y. Im, and M. Chiang. Tube: time-dependent pricing for mobile data. *ACM SIGCOMM Computer Communication Review*, 42(4):247–258, 2012.

[58] C. Higson and D. Waltho. Valuing information as an asset, 2009.

[59] R. A. Howard. Information value theory. *Systems Science and Cybernetics, IEEE Transactions on*, 2(1):22–26, 1966.

[60] F. Huang and Y. Han. Price discovery, competition and market mechanism design. *Asian Social Science*, 4(6):122, 2009.

[61] S. Jain and P. Kannan. Pricing of information products on online servers: Issues, models, and analysis. *Management Science*, 48(9):1123–1142, 2002.

[62] J. Jaisingh, J. Barron, S. Mehta, and A. Chaturvedi. Privacy and pricing personal information. *European Journal of Operational Research*, 187(3):857–870, 2008.

[63] N. Jentsch, S. Preibusch, and A. Harasser. Study on monetising privacy: An economic model for pricing personal information. *ENISA, Feb*, 2012.

[64] H. Jin, M. Xiong, and S. Wu. Information value evaluation model for ilm. In *Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing, 2008. SNPD'08. Ninth ACIS International Conference on*, pages 543–548. IEEE, 2008.

[65] A. Karr, A. Oganian, J. Reiter, and M.-J. Woo. New measures of data utility. In *Workshop Manuscripts of Data Confidentiality, A Working Group in National Defense and Homeland Security. Available at http://sisla06. samsi. info/ndhs/dc/Papers/NewDataUtility-01-10-06. pdf*, 2006.

[66] J. M. Keisler, Z. A. Collier, E. Chu, N. Sinatra, and I. Linkov. Value of information analysis: the state of application. *Environment Systems and Decisions*, 34(1):3–23, 2014.

[67] R. Kesavamoorthy, D. Arunshunmugam, and L. Thangamariappan. Solving traveling salesman problem by modified intelligent water drop algorithm. In *International Conference on Emerging Technology Trends (ICETT). Proceedings published by International Journal of Computer Applications (IJCA)*, volume 2, pages 18–23, 2011.

[68] T. Kiemes, D. Oberle, and F. Novelli. Towards a reusable and executable pricing model in the internet of services. In *Proceedings of the 12th International Conference on Information Integration and Web-based Applications & Services*, pages 722–729. ACM, 2010.

[69] K.-j. Kim and I. Han. Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index. *Expert systems with applications*, 19(2):125–132, 2000.

[70] K.-S. Kim. Measures of data utility for complex survey data.

[71] J. Kleinberg, C. H. Papadimitriou, and P. Raghavan. On the value of private information. In *Proceedings of the 8th conference on Theoretical aspects of rationality and knowledge*, pages 249–257. Morgan Kaufmann Publishers Inc., 2001.

[72] P. Koutris, P. Upadhyaya, M. Balazinska, B. Howe, and D. Suciu. Query-based data pricing. In *Proceedings of the 31st symposium on Principles of Database Systems*, pages 167–178. ACM, 2012.

[73] A. Krause and E. Horvitz. A utility-theoretic approach to privacy in online services. *arXiv preprint arXiv:1401.3859*, 2014.

[74] R. Kumar, D. Sharma, and A. Sadu. A hybrid multi-agent based particle swarm optimization algorithm for economic power dispatch. *International Journal of Electrical Power & Energy Systems*, 33(1):115–123, 2011.

[75] K. C. Laudon. Markets and privacy. *Commun. ACM*, 39(9):92–104, Sept. 1996.

[76] D.-W.-I. S. Lehmann and P. Buxmann. Pricing strategies of software vendors. *Business & Information Systems Engineering*, 1(6):452–462, 2009.

[77] S. Lehmann, T. Draisbach, P. Buxmann, and P. Drsam. *Pricing of software as a servicean empirical study in view of the economics of information theory*, pages 1–14. Springer, 2012.

[78] E. Levin, R. Pieraccini, and W. Eckert. Using markov decision process for learning dialogue strategies. In *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, volume 1, pages 201–204. IEEE, 1998.

[79] C. Li, D. Y. Li, G. Miklau, and D. Suciu. A theory of pricing private data. In *Proceedings of the 16th International Conference on Database Theory*, pages 33–44. ACM, 2013.

[80] X. Liang, X. Li, R. Lu, X. Lin, and X. Shen. Udp: Usage-based dynamic pricing with privacy preservation for smart grid. *Smart Grid, IEEE Transactions on*, 4(1):141–150, 2013.

[81] J. Lin, S. Amini, J. I. Hong, N. Sadeh, J. Lindqvist, and J. Zhang. Expectation and purpose: understanding users' mental models of mobile app privacy through crowdsourcing. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pages 501–510. ACM, 2012.

[82] Q. Liu, T. Luo, R. Tang, and S. Bressan. An efficient and truthful pricing mechanism for team formation in crowdsourcing markets. In *Communications (ICC), 2015 IEEE International Conference on*, pages 567–572. IEEE, 2015.

[83] Y. Liu, K. P. Gummadi, B. Krishnamurthy, and A. Mislove. Analyzing facebook privacy settings: user expectations vs. reality. In *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*, pages 61–70. ACM, 2011.

[84] Y. Liu and P. Yuan. Free or free and fee: Pricing strategy of information goods on mobile internet. In *Electronic Commerce and Security (ISECS), 2010 Third International Symposium on*, pages 321–323. IEEE, 2010.

[85] A. Lobianco, A. Dragicevic, and A. Leblois. Forest planning and productivity-risk trade-off through the markowitz mean-variance model. 2015.

[86] L. Longpre and V. Kreinovich. How to measure loss of privacy. 2006.

[87] D. Ma and J. Huang. The pricing model of cloud computing services. In *Proceedings of the 14th Annual International Conference on Electronic Commerce*, pages 263–269. ACM, 2012.

[88] M. Macauley and R. Laxminarayan. The value of information:methodological frontiers and new applications for realizing social benefitworkshop. *Space Policy*, 26(4):249–251, 2010.

[89] M. Macías and J. Guitart. A genetic model for pricing in cloud computing markets. In *Proceedings of the 2011 ACM Symposium on Applied Computing*, pages 113–118. ACM, 2011.

[90] M. Madden, A. Lenhart, S. Cortesi, and U. Gasser. Teens and mobile apps privacy. *Pew Internet and American Life Project*, 2013.

[91] C. Maina. Valuing information in an information age: The price model and the emerging information divide among individuals, societies, and nations. *Canadian Journal of Information and Library Science*, 27(3):139, 2003.

[92] R. Mansini, W. Ogryczak, and W. G. Speranza. Portfolio optimization. In *Linear and Mixed Integer Programming for Portfolio Optimization*, pages 1–18. Springer, 2015.

[93] F. G. Marmol, C. Sorge, O. Ugus, and G. M. Pérez. Do not snoop my habits: preserving privacy in the smart grid. *Communications Magazine, IEEE*, 50(5):166–172, 2012.

[94] C. Marsh. Introduction to continuous entropy, 2013.

[95] F. McSherry and K. Talwar. Mechanism design via differential privacy. In *Foundations of Computer Science, 2007. FOCS'07. 48th Annual IEEE Symposium on*, pages 94–103. IEEE, 2007.

[96] L. Mei, W. Li, and K. Nie. *Pricing Decision Analysis for Information Services of the Internet of Things Based on Stackelberg Game*, pages 1097–1104. Springer, 2013.

[97] Y. Meng, O. Kazeem, and J. Muller. A swarm intelligence based coordination algorithm for distributed multi-agent systems. In *Integration of Knowledge Intensive Multi-Agent Systems, 2007. KIMAS 2007. International Conference on*, pages 294–299. IEEE, 2007.

[98] J. Mikians, L. Gyarmati, V. Erramilli, and N. Laoutaris. Detecting price and search discrimination on the internet. In *Proceedings of the 11th ACM Workshop on Hot Topics in Networks*, pages 79–84. ACM, 2012.

[99] J. Mikians, L. Gyarmati, V. Erramilli, and N. Laoutaris. Crowd-assisted search for price discrimination in e-commerce: First results. In *Proceedings of the ninth ACM conference on Emerging networking experiments and technologies*, pages 1–6. ACM, 2013.

[100] K. Mivule. Utilizing noise addition for data privacy, an overview. *arXiv preprint arXiv:1309.3958*, 2013.

[101] M. Murthy, H. Sanjay, and J. Ashwini. Pricing models and pricing schemes of iaas providers: a comparison study. In *Proceedings of the International Conference on Advances in Computing, Communications and Informatics*, pages 143–147. ACM, 2012.

[102] M. Nagarajan, K. Baid, A. Sheth, and S. Wang. Monetizing user activity on social networks-challenges and experiences. In *Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology-Volume 01*, pages 92–99. IEEE Computer Society, 2009.

[103] T. Nagle. Economic foundations for pricing. *Journal of Business*, pages S3–S26, 1984.

[104] M. Naldi and L. Mastroeni. Cloud storage pricing: a comparison of current practices. In *Proceedings of the 2013 international workshop on Hot topics in cloud services*, pages 27–34. ACM, 2013.

[105] K. Nissim, C. Orlandi, and R. Smorodinsky. Privacy-aware mechanism design. In *Proceedings of the 13th ACM Conference on Electronic Commerce*, pages 774–789. ACM, 2012.

[106] Oecd. Exploring the economics of personal data: A survey of methodologies for measuring monetary value. OECD Digital Economy Papers 220, OECD Publishing, 2013.

[107] D. L. Olive Huang. How organizations can monetize customer data. Technical report, Gartner, 2014.

[108] P. Racherla, M. J. Keith, and J. S. Babb Jr. Pay-what-you-want pricing for mobile applications: The effect of social information and privacy assurances. 2011.

[109] A. P. Rajan et al. Evolution of cloud storage as cloud computing infrastructure service. *arXiv preprint arXiv:1308.1303*, 2013.

[110] D. Reed and E. H. Chi. Online privacy; replicating research results. *Communications of the ACM*, 55(10):8–9, 2012.

[111] C. Riederer, V. Erramilli, A. Chaintreau, B. Krishnamurthy, and P. Rodriguez. For sale: your data: by: you. In *Proceedings of the 10th ACM WORKSHOP on Hot Topics in Networks*, page 13. ACM, 2011.

[112] M. Sajko, K. Rabuzin, and M. Baa. How to calculate information value for effective security risk assessment. *Journal of Information and Organizational Sciences*, 30(2):263–278, 2006.

[113] P. Samadi, A.-H. Mohsenian-Rad, R. Schober, V. W. Wong, and J. Jatskevich. Optimal real-time pricing algorithm based on utility maximization for smart grid. In *Smart Grid Communications (SmartGridComm), 2010 First IEEE International Conference on*, pages 415–420. IEEE, 2010.

[114] P. Samadi, R. Schober, and V. W. Wong. Optimal energy consumption scheduling using mechanism design for the future smart grid. In *Smart Grid Communications (SmartGridComm), 2011 IEEE International Conference on*, pages 369–374. IEEE, 2011.

[115] H. Shah-Hosseini. Problem solving by intelligent water drops. In *Evolutionary Computation, 2007. CEC 2007. IEEE Congress on*, pages 3226–3231. IEEE, 2007.

[116] H. Shah-Hosseini. Intelligent water drops algorithm: A new optimization method for solving the multiple knapsack problem. *International Journal of Intelligent Computing and Cybernetics*, Vol. 1 Iss: 2:193 – 212, 2008.

[117] C. E. Shannon. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1):3–55, 2001.

[118] W. F. Sharpe. Mutual fund performance. *The Journal of business*, 39(1):119–138, 1966.

[119] S. Shenker, D. Clark, D. Estrin, and S. Herzog. Pricing in computer networks: Reshaping the research agenda. *Telecommunications Policy*, 20(3):183–201, 1996.

[120] R. Siciliano. Data brokers: What are they; how to get control of your name, April 21 2014. Retrieved June 20, 2015.

[121] N. Singer. Data broker is charged with selling consumers' financial details to 'fraudsters', 23 December 2014. Retrieved June 20, 2015.

[122] E. G. Smit, G. Van Noort, and H. A. Voorveld. Understanding online behavioural advertising: User knowledge, privacy concerns and online coping behaviour in europe. *Computers in Human Behavior*, 32:15–22, 2014.

[123] Y.-Y. Sohn. Asymmetry in pricing information goods. In *Frontiers of Broadband, Electronic and Mobile Commerce*, pages 181–192. Springer, 2004.

[124] H. Song, C.-C. Liu, J. Lawarrée, and R. W. Dahlgren. Optimal electricity supply bidding by markov decision process. *Power Systems, IEEE Transactions on*, 15(2):618–624, 2000.

[125] A. C. Squicciarini, M. Shehab, and J. Wede. Privacy policies for shared content in social network sites. *The VLDB JournalThe International Journal on Very Large Data Bases*, 19(6):777–796, 2010.

[126] F. Stahl and G. Vossen. High quality information provisioning and data pricing. In *Data Engineering Workshops (ICDEW), 2013 IEEE 29th International Conference on*, pages 290–293. IEEE, 2013.

[127] G. J. Stigler. The economics of information. *The journal of political economy*, pages 213–225, 1961.

[128] A. Sullivan and S. M. Sheffrin. Economics: Principles in action. upper saddle river, new jersey 07458: Pearson prentice hall. 2003.

[129] Z. Sun and L. Li. Potential capability estimation for real time electricity demand response of sustainable manufacturing systems using markov decision process. *Journal of Cleaner Production*, 65:184–193, 2014.

[130] K. P. Sycara. Multiagent systems. *AI magazine*, 19(2):79, 1998.

[131] C. R. Taylor. Consumer privacy and the market for customer information. *RAND Journal of Economics*, pages 631–650, 2004.

[132] D. Teodorovic. Transport modeling by multi-agent systems: a swarm intelligence approach. *Transportation Planning and Technology*, 26(4):289–312, 2003.

[133] E. Toch, Y. Wang, and L. F. Cranor. Personalization and privacy: a survey of privacy risks and remedies in personalization-based systems. *User Modeling and User-Adapted Interaction*, 22(1-2):203–220, 2012.

[134] M. C. Tschantz and A. V. Nori. Measuring the loss of privacy from statistics. In *Proceedings of the 1st Workshop on Quantitative Analysis of Software (QA09), Technical Report UCB/EECS-2009-93, Electrical Engineering and Computer Sciences, University of California at Berkeley*, pages 27–36, 2009.

[135] P. Upadhyaya, M. Unutzer, M. Balazinska, D. Suciu, and H. Hacigumus. Affordable analytics on expensive data. In *Proceedings of the First International Workshop on Bringing the Value of Big Data to Users (Data4U 2014)*, page 19. ACM, 2014.

[136] H. R. Varian. Pricing information goods, 1995.

[137] H. Von StackelberG, D. Bazin, R. Hill, and L. Urch. *Market structure and equilibrium*. Springer, 2010.

[138] J. Waters. Pricing information goods with piracy and heterogeneous consumers. *Nottingham University Business School Research Paper*, (2013-07), 2013.

[139] L. Wathieu and A. A. Friedman. An empirical approach to understanding privacy valuation. *HBS Marketing Research Paper*, (07-075), 2007.

[140] T. Weber. Price theory in economics. 2012.

[141] T. A. Weber and D. C. Croson. Selling less information for more: garbling with benefits. *Economics Letters*, 83(2):165–171, 2004.

[142] A. Whitmore, D. F. Andersen, J. Zhang, and L. F. Luna-Reyes. A policy framework for evaluating full information product pricing (fipp) regimes. In *Proceedings of the 11th annual international digital government research conference on public administration online: Challenges and opportunities*, pages 233–234. Digital Government Society of North America, 2010.

[143] Wikipedia. Value (economics) — wikipedia, the free encyclopedia, 2014. [Online; accessed 19-February-2014].

[144] K.-W. Wu, S. Y. Huang, D. C. Yen, and I. Popova. The effect of online privacy policy on consumer privacy concern and trust. *Computers in human behavior*, 28(3):889–897, 2012.

[145] H. Xu and B. Li. A study of pricing for cloud resources. *ACM SIGMETRICS Performance Evaluation Review*, 40(4):3–12, 2013.

[146] F. Yokota and K. M. Thompson. Value of information literature analysis: a review of applications in health risk management. *Medical Decision Making*, 24(3):287–298, 2004.

[147] M. Yu, K. Yang, L. Wei, and J. Sun. Practical private information retrieval supporting keyword search in the cloud. In *Wireless Communications and Signal Processing (WCSP), 2014 Sixth International Conference on*, pages 1–6. IEEE, 2014.

[148] X. Zhang, Z. Yang, Z. Zhou, H. Cai, L. Chen, and X. Li. Free market of crowdsourcing: Incentive mechanism design for mobile sensing. *Parallel and Distributed Systems, IEEE Transactions on*, 25(12):3190–3200, 2014.

[149] B. Zhao, C. Guo, and Y. Cao. A multiagent-based particle swarm optimization approach for optimal reactive power dispatch. *Power Systems, IEEE Transactions on*, 20(2):1070–1078, 2005.

[150] D. Zhao, X.-Y. Li, and H. Ma. How to crowdsource tasks truthfully without sacrificing utility: Online incentive mechanisms with budget constraint. In *INFOCOM, 2014 Proceedings IEEE*, pages 1213–1221. IEEE, 2014.

[151] W. Zhao and Y.-S. Zheng. Optimal dynamic pricing for perishable assets with nonhomogeneous demand. *Management science*, 46(3):375–388, 2000.