경영학석사학위논문

# The Role of a Messaging System in Initial Online Freelance Market

# 초기 온라인 프리랜서 시장 수요 분석 및 메시지 시스템의 효과 분석

**2016 년 8 월**

서울대학교 대학원

경영학과 경영학 전공

신 민 규

# Abstract

Online freelance market is made up of two major players: Buyers and Sellers. After joining the website of an online freelance company, a buyer would post a project with the detailed description about the work or the explanation about an ideal seller. After a buyer finishes posting her projects, potential sellers take part in the competition to win the business deal. The whole procedure of a transaction takes a reverse auction process. In other words, sellers write out a bid which includes “bid price” and “bid period”. (bid price refers to how much sellers would charge for carrying out the project and bid period refers to the expected time required to complete the project.) After examining all the bid-specific information, the buyer would choose the winner of the auction. In this process, market players, especially buyers, may encounter serious information asymmetry problem.

The seller who suggests a bid price has more information about the final output of the project than a buyer has. That being said, sellers who have more inside information also have a chance to give a bid in their favor. The inside information would be the true capability to carry out the project and/or the willingness to put effort into the project. An important issue in this market would be the possible

imbalance of the information. In economic context, the “lemon” problem from information asymmetry may drive out market players who would be so diligent, sincere, and competent that they could contribute market-settlement. The informational asymmetry may result in a market that retains only the players having low level of capability, accelerating more exit of market players. This vicious circle is a matter of life and death to the operator of the online freelance service. So it is very important for the online freelance operator to relieve the information asymmetry problem.

Online freelance companies can consider introducing a messaging system within the online auction process. Through the messaging system, sellers (freelancers) have an opportunity to appeal themselves about their willingness to do a project and to provide the information about the ability they have for conducting a project. Therefore, adopting the messaging system seems to be an appropriate solution to relieve the information asymmetry problem that online freelance companies have. But the actual effect of message in the decision process of buyers (clients) has not been explored much.

This study builds a structural dynamic model of utility maximizing behaviors of buyers and estimates the model using a large dataset obtained from a Korean online freelance website. Buyers are modeled

to be dynamic utility maximizer in the sense that a buyer, in each time period, has to make a decision on whether to select a bid or to wait for another bid. If more promising bid is expected to come, the buyer may wait for future bid. Otherwise, the buyer may select a bid among the currently available ones. In order to incorporate such feature in buyer behaviors, this study utilizes the finite horizon dynamic programming approach in modeling explicitly the buyer behaviors.

The data set comes from a domestic online freelance marketplace. An interesting feature of the site is that sellers can send messages to buyers and the message data are stored in the website database. This study explores the role of the message system of the site in reducing the information asymmetry problem. In terms of estimation, we first estimated a model that does not incorporate the message into the utility function. In the model without the message data, buyers turned out insensitive to bid price. Note that if one does not control for the product quality in the demand function where the quality indeed affects the demand and at the same time is correlated with price, the price effect could be underestimated in the model without the quality. This typical omitted variable bias may kick in the online freelance market problem if the bid price indeed reflects the true capability of the sellers. That is, the sellers with a higher level of capability may

post higher prices and the buyers may take such capabilities into consideration when marking decision. If that is the case, the quality of bid is observed by the decision making buyers but not by researchers. To control for the unobserved quality, we include the message data into out model. If the messages from the sellers contain the information that buyers can utilize to assess the capability of the bids, the model with the message data can better identify the true effect of bid price. In order to incorporate of the contents of the message into the model, we transformed the text information in the message dataset into numeric data via a topic modeling approach. In the model with the message data, the buyers are estimated to be more price sensitive than in the model without the message information.

***Keywords:*** Structural Approach; Dynamic Programming; Text Data; Topic Modeling; Online Freelance Market

***Student Number:*** 2014-20435

# Table of Contents

# 1. Introduction

1.1 Introduction to online freelance market

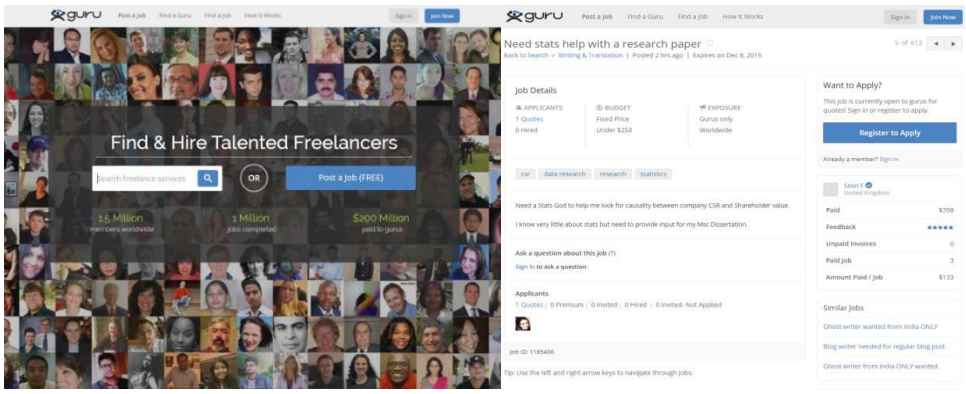
As technology is getting developed, totally new industries emerge in various fields. Among them, it is remarkable that the P2P mediating market accomplishes enormous size of growth. To be specific, online freelance market is worthy of the center of attention. This market functions as an intermediary agent to connect two players : buyers and sellers. Buyers are the ones who ask an outside expert, or seller, to take their project and sellers are the ones who want to earn money by selling their labor. Companies in this market make a profit by charging a given portion of the selected bid as a fee. For example, let's assume that a local government wants to open a new homepage to advertise its local festival such as a traditional art festival. Usually, there are not enough experts to take the role of the web-page production in the local government. This makes the director in charge of the festival depend on the outsourcing. In the past, it would be possible, at best, to employ an acquaintance in local community. But now, with the help of the online mediating company, the director can select the most effective offer among the nationwide expert pools. Enjoying the bigger size of supply side (sellers), buyers can compare lots of candidates and choose the optimal one. To activate this market, the online company should give enough trusts on the quality of the online service to both sides of participants.

This online freelance market belongs to the two-sided market in which preoccupancy or staying ahead of the game is very important for competitive advantage. Lots of previous studies in industrial organization pointed out this crucial network effect. Like what previous papers said, it is vital to penetrate the market as quickly as possible. But taking root in online business environment is rather heavy going. This is because it is not easy for companies to lay a foundation of trusts to make market

players like sellers and buyers participate this market. International companies such as Elance overcame the initial obstacle very successfully and have been on the up and up. In South Korea, nowadays, there are companies to follow this golden path and pioneer a new market. One of the leading companies has grown about 30% every quarter this year and attracted a big money for investment. However, this market is not full of prosperous and flourishing future. Its inside story tells a severe problem from which this initial market suffers. Before diving into this problems, it is necessary to take a closer look at this market structure.

Online freelance market is made up of two major players like what I explained above. After joining the website of an online freelance company, buyers could post a project with detailed description about the work or explanation about an ideal seller. Figure 1 shows a good example of this procedure.

<Figure 1> Example of an online freelance company : Guru




After buyers finish posting their projects, sellers take part in the competition to win the business deal. What's more interesting is that the whole procedure of a transaction takes a reverse auction process. In other words, sellers write out a bid which includes "bid price" and "bid period". (bid price refers to how much sellers would charge for carrying out the project and bid period means how long the project is going to take.) After examining all the bid-specific information, buyers would choose who will be the winner of the auction or who will be responsible for the project. In this process, market players, especially buyers, may encounter serious

information asymmetry problem. The entity, seller, who suggests a bid price is the same one who has more information about the final output of the project. That being said, sellers who have more inside information also have a chance to give a bid in their favor. This is because the inside information can be the true capacity to carry out the project or the willingness to put meaningful effort into the project. What counts is that the imbalance of the information is quite severe. In economic context, the "lemon" problem from information asymmetry drives out market players who would be so diligent, sincere, and competent that they could contribute market-settlement. The problem leaves only the players having low level of capability, accelerating more exit of market players. This vicious circle is a matter of life and death to the company. So it is the most important task for the company to cut the circle by attenuating the information asymmetry problem. To handle the problem, lots of remedies have been discussed and applied in the market.

The most effective and intuitive solution is to construct a reviewing process or to build a reputation system. If the information containing an evaluation on the output of the service is accumulated and is made open to market players, the problem can be mitigated. Moreover, it is a well-suited method for this kind of online market in that the whole procedure can be simply added to the existing online business model. But there are necessary conditions for this remedy to be a universal key. First, to make players utilize the reputation meaningfully, companies should accumulate enough amount of reputation data in advance of market players' getting into trouble by information asymmetry. Unfortunately, reputation building needs quite long time during which market players might be disappointed at the quality of service and be turned into callous customers. That is to say, there is not enough time for a young company to wait for the stack of reputation because it should prove its promising performance based on its current status quo to attract investment from venture capital companies. And network effect allows only a few companies to enter the realm of

dreams. To be the winner that takes it all, the company has to be hurry by all means. So it may need another solution especially at the early period.

The most immediate antidote other than a reputation system is to arrange an offline meeting between buyers and sellers. This pill would be another prescription for dealing with the information asymmetry problem in online freelance market. In this way, sellers who are selected for offline meeting have a chance to share their own inside information with buyers. Unfortunately, it is possible only when market size is small. This is due to the fact that this remedy increases cost for sure, but does not guarantee revenue increase. Fixing up and setting up the offline meeting needs company to send its employee to that place. This cost occurs whenever the offline meeting is arranged. But the final contract may not be made even after the offline meeting because buyers have the option not to choose a seller even after an offline meeting. So when market size becomes bigger, the profit of the company will be encumbered by this makeshift.

Last but not least, companies can consider introducing a messaging system within the online auction process. Through the messaging system, sellers have an opportunity to appeal themselves about their willingness to do a project and to provide the information about the ability they have for conducting a project. Although online freelance companies provide sellers' profile such as their career, skills, and education in the website, the information in the profile is too limited for buyers to judge lots of sellers. Compared to the basic profile, the messaging system can generate a space where sellers can provide detailed information like the reason why their price in the bid is so high or why they need lots of time to complete the project. So it looks more effective for both side of players. In addition, the messaging system does not need so much cost to construct the system. Therefore, adopting the messaging system seems to be the most appropriate solution for an initial online freelance company. But actual effect of message in the decision process of buyers

has not been unearthed until this paper. This study utilizes a message data with the help of Topic Modeling and includes the feature of message data in the framework of dynamic programming. In this way, I can quantify the value of message as well as other bid-related features such as price. I discover that buyers could be insensitive to price especially in the initial online market which is full of uncertainty. This might be because quality information was mingled with price. After controlling the message data that contained the quality information, I can observe a more plausible price coefficient.

## 1.2 Introduction to this paper

At this part, I briefly explain the model, the estimation, and the result. The set of control variable, or choice variable, for buyer side exhaustively consists of three components. During decision process, buyers can choose one of these three options: "waiting", "canceling the project", and "choosing one of the bids arrived". Among the three options, "waiting" behavior instills the need for researchers to concentrate on the dynamic aspects of buyers' behavior. By waiting at time t, the buyer would have ambivalent effects on her utility. When a decision maker waits at time t, she can get a better bid which is more suited to her project. Sometimes, the bid with low price and short period might come from future sellers who are more skilled in the project-relevant area. This is the bright side of choosing a "waiting" option. But she would have to sacrifice the current utility from starting the project at time t or from canceling the project. This complex effect with regard to decision time transforms the decision maker's choice problem into the optimal stopping problem which is actively studied in the marketing and economics field.

On top of that, researchers need to consider the unobserved heterogeneity embedded in buyers. Let's assume that there are two types of buyers. These buyers are thought to be identical on the perspective of observed covariate data, or independent

variables, including project size, number of arrived bids, characteristics of those bids, and so on. In this situation, buyers would be expected to behave in the same way without the unobserved heterogeneity. (More specifically, in discrete choice model, this situation is equivalent to that there is no heterogeneity to explain the discrepancy between models and data other than idiosyncratic error term.) If researchers are so lucky that they can collect all the possible source of heterogeneity of buyers, then they might not worry about this problem. Unfortunataely, these environments rarely happen. What's worse is that unobserved heterogeneity becomes more difficult to combine when it is mingled with dynamic problem. Yoganarasimhan (2013) refers to the problem as *dynamic selection*. One of the explanations about dynamic selection is that depending on buyers, outside alternatives are so different. Some buyers are likely to be more prepared to do their project for themselves, which is not revealed in data. They are not likely to wait for a long time, immediately canceling their projects. Researchers should consider how to combine this unobserved heterogeneity within dynamic programming. Not combining the heterogeneity may change other parameters into biased estimators.

I have to set a choice model taking the unobserved heterogeneity into the account. In addition to the unobserved state space, there is a traditional difficulty in dealing with the curse of dimensionality when decision maker considers numerous state variables in decision process. If that is the case, it may be more appropriate to use Conditional Choice Probability approach (Hotz and Miller 1993). This two-step method allows the researchers to directly estimate the value function from the data. Compared to nested fixed-point algorithms (Rust 1987), this method (CCP henceforth) can work well in a high dimensional problem when dataset is rich. But the critical weakness of CCP is a difficulty in dealing with an unobserved heterogeneity. This is because the unobserved variable cannot be measured in dataset, making the substitution of value function with CCP impossible. To make up for the

stumbling block, researchers developed CCP with unobserved heterogeneity (Arcidiacono and Miller 2011). This method takes the ways to handle missing value like EM-algorithm (Dempster et al. 1977). In the marketing field, using the newly developed estimation technique, many researchers have been illuminating various kinds of decision problem. Chung et al. (2013) utilize the method to analyze the relationship between a bonus-based compensation and workers' productivity. Among considerable studies, Yoganarasimhan (2013) is similar to my paper except the point that I focus on evaluating a messaging system.

Following the estimation method explained above, I estimate the parameters of buyers' utility and discover three notable points. First of all, what's interesting is that proposed price contained in a bid could function as the cue for quality of a bid such as the experience of sellers. The result touches a similar spirit of what prior researches suggested in the name of *hedonic price*. Second, the result shows that buyers can be divided into two segments in the proportion of 2:8. Smaller group is expected to feel greater utility to choose the bid, compared to the majority group. The ratio seems quite different from the prior result of Yoganarasimhan (2013) which suggested the ratio of 3:7. This difference may be due to the uncertainty of initial online market. Thirdly, I find out that repetitive buyers, who post more than one project over different periods, show a higher tendency to choose a bid. Especially, the first result regarding the issue of quality control has a lot to discuss. At first, I estimated the demand function of buyers, not including the message data. That result revealed that buyers seem insensitive to price in a bid, which sounded very weird. But, if I did not control the quality part which may be positively correlated with bid price, the coefficient of price would reflect the mingled effect, not partial effect of price. This problem might be thought of as a typical omitted variable problem in econometrics context other than the point that the issue should be understood in the context of structural estimation. To control the quality part, I combine the message

data with the existing bid-related data. Message data has the information that buyers may utilize to guess the quality of each bid offer. By transforming the message dataset into a more manageable form and controlling the message data variation, I can discover that the effect of price would seem to be more plausible.

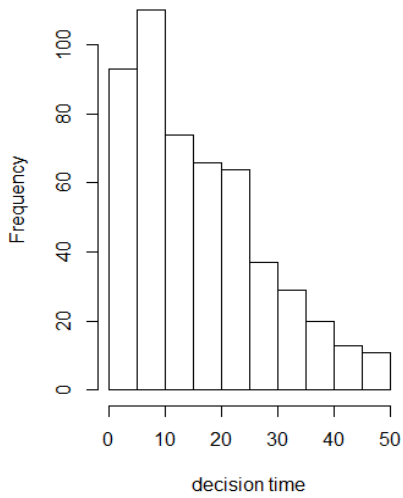
## 2. Data Overview

One of venture companies doing online freelance business in South Korea provided me with the dataset from October 2013 to April 2015. The dataset contains all the information about buyers' behavior and sellers' behavior, including the message data generated by sellers. For instance, let's assume that there are a buyer and four sellers in the online company. Company stores the profile of these customers, which was generated at their enrollment time. If the buyer posts a project, then the dataset gets bigger, including the detailed information about project price which is set by the buyer and project period also described by her. In addition, all the bid-specific information from each bid of sellers is also accumulated, such as price, period. If a seller uses the messaging system, the message from the seller to buyer is also piled up. This big chunk of dataset is offered from the Korean company. Unfortunately, some crucial parts of dataset from October 2013 to August 2014 was erased, making the data in this period useless. So the remaining part of dataset was used for structural estimation. Total number of remaining projects is 517 and total number of bids which sellers proposed is 6347. Average time to reach the terminal decision that buyers stop their optimal stopping problem is about 16 days. Figure 2 shows the distribution of decision time. And Figure 3 shows the increasing pattern of a cancel ratio as buyers defer the final decision.[①] This pattern almost exactly corresponds to the decreasing

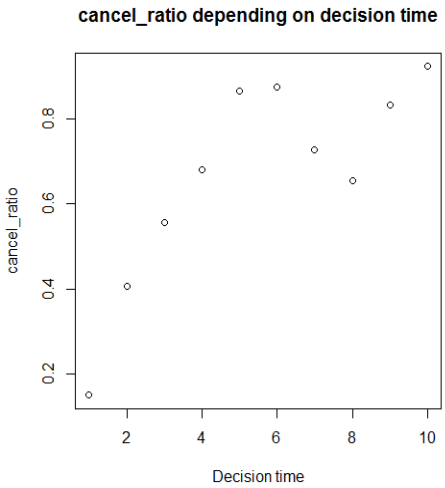
① The scale of x-axis is 5 days ( e.g. decision time 2 means he or she made decision after 10 days )

pattern of selection ratio in Yoganarasimhan (2013). So the pattern detects that there is “dynamic selection” problem in my dataset like the previous research.

<Figure 2> Decision time distribution




<Figure 3> Increasing cancel ratio pattern




To explain the dynamic selection over time, the previous study claimed the necessity to combine the unobserved heterogeneity. According to the paper, buyers who showed high propensity to wait were the self-selected group by the unobserved heterogeneity. To be specific, surviving decision makers who postpone the final decision might have a lower utility for choosing a bid among arrivals, which is related to the decreasing pattern of “select” ratio upon decision time. (Or it can be interpreted as the increasing pattern of “cancel” ratio upon decision time as Figure 3) So, not combining the unobserved heterogeneity can bring the serious biased estimation for the effect of explanatory variables.

Table 1 and Figure 4 shows more detailed information about heterogeneity between buyers. Average project price (guideline price which was suggested by buyers) is about 9 million won and average project period (guideline period for the project to

be finished) is about 40 days. Guideline values for projects could function as the accelerator to concentrate the bid value, like bid price, on that guideline value. Actual distribution of bid value looks very similar to normal distribution, which helped me to use the parametric approach for transition probability distribution unlike the previous research. Table 1 brightens some interesting facts about the heterogeneity between buyers. Second column of detailed information section in table 1 is about how bid price varies with the project price. For auctions that have lower project price, the mean of normalized bid price (bid price divided by project price) is relatively high. As project size becomes larger and project price gets bigger, the average of normalized bid price arrived at the auctions come to decrease steadily, converging to guideline value, project price.

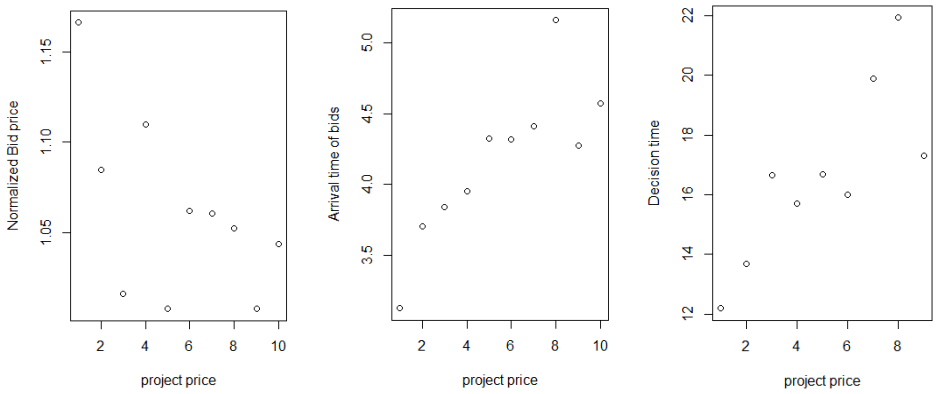
<Table 1> Summary Statistics of projects

| Summary statistics of projects | | | | | | |
|---|---|---|---|---|---|---|
| | Min | Max | Mean | 1Q | 2Q | 3Q |
| Project price (1000 won) | 100 | 100000 | 8761 | 2000 | 5000 | 10000 |
| Project period (days) | 1 | 180 | 37.91 | 15 | 30 | 60 |
| Detailed information for each project group | | | | | | |
| Project Group (project price) | Number Of bids | Normalized Bid price (mean) | Arrival Time of bids (mean) | Project Subfield (ratio) | Decision Time (days) | Cancel Ratio |
| ~10% | 608 | 1.1661 | 3.1217 | 0.6806 | 12.1805 | 0.4861 |
| 10%~20% | 352 | 1.0845 | 3.7017 | 0.3243 | 13.6757 | 0.7027 |
| 20%~30% | 477 | 1.0162 | 3.8365 | 0.4255 | 16.6596 | 0.4255 |
| 30%~40% | 683 | 1.1099 | 3.9502 | 0.4000 | 15.7091 | 0.5091 |
| 40%~50% | 859 | 1.0077 | 4.3248 | 0.1410 | 16.6795 | 0.6026 |
| 50%~60% | 327 | 1.0620 | 4.3150 | 0.1250 | 16.0000 | 0.6563 |
| 60%~70% | 1059 | 1.0606 | 4.4089 | 0.1084 | 19.9036 | 0.6506 |
| 70%~80% | 150 | 1.0522 | 5.1600 | 0.1429 | 21.9286 | 0.5714 |
| 80%~90% | 914 | 1.0081 | 4.2713 | 0.0526 | 17.2983 | 0.6140 |
| 90%~ | 859 | 1.0436 | 4.5716 | 0.0476 | 19.5238 | 0.5238 |

In contrast, the average of arrival time of bids increases as the size of project gets bigger. These two different aspects of bids are quite clear, as you can unmistakably

catch the pattern in the first picture of Figure 4. Sellers might have some interesting behavior patterns based on their own optimization. But that topic is beyond the scope of my research. About this issue, Yoganarasimhan (2015) incorporates the optimization of sellers' side. The last picture of Figure 4 describes the pattern of buyers' behavior. As project size increases, they tried to spend more time to make final decision. But the descriptive statistics is just the simple and one dimensional picture of behavior pattern, not explaining the multi-dimension aspects of behavior. To do that, researchers should consider various variables simultaneously in the model. From now on, I will explain the variables that I take account of.

<Figure 4> Plots of Table 1

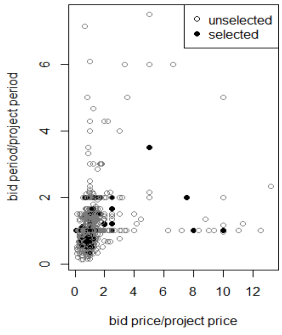



2.1. Structured Data

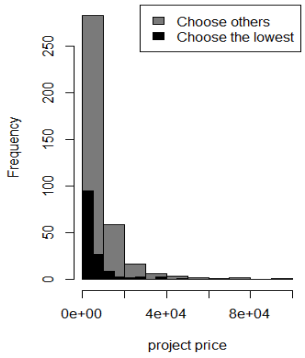
I utilized almost the same dataset with Yoganarasimhan (2013) other than unstructured data like message. Structured Data consists of two sources: buyers and sellers. Figure 5 shows the core part of data from sellers. Sellers generated the information about bid price (How much), bid period (How long), when participating in auctions. On top of that, the company assigned the "description index" to each seller depending on the degree of completion of seller's profile. Buyers were allowed to access the index, possibly making use of that when they decided the final winner. So I had to combine the index with bid-specific dataset. (It would be more desirable

to combine the information about sellers' profile, if possible. But the information looks too messy to be combined with other structured data, making me use proxy variable like description index.) These characteristics of sellers, or bids, may be related to the decision of buyers. Figure 5 shows the relationship between bid price and bid period. It seems quite natural to say that buyers would prefer low bid price and short bid period. What Figure 5 describes is, however, that there is also a dispersion in the buyers' selection. Figure 6 also confirms that there is a hidden criterion which buyers might use other than price. If buyers had used the only criterion like price, they would have chosen the lowest priced bid offer. But it isn't matched with Figure 6. Figure 6 also shows that the portion of selecting the lowest bid offer slowly decreases as project price (a guideline price suggested by buyers) goes up. One interpretation is that as the willingness to pay for buyers (project price) increases, they might check all the information they could use meticulously, other than price. This guess would lead me to combine the message data which has been out of use.

<Figure 5> Relationship between bid price and bid period (normalized by project size)




<Figure 6> Selection of the lowest priced bid



## 2.2 Unstructured Data

In a messaging system, sellers had an opportunity to convey the inside information about their capability to carry out the project, or their willingness to accomplish the project. After a long letter was received to a buyer, the buyer would infer the potential quality that the seller had. Message was written by Korean sellers in Korean, making me solve a stack of problem in natural language processing field. Problem seemed to be more severe because of the complex grammar structure in Korean, hindering me from utilizing the developed method in machine learning field. So I translated the message data from Korean to English by Google Translation Toolkit. Figure 7 reveals one example of the results of translation. Grammar structure and connotation might be destroyed in this simple data imputation. Even with the defect, I could apply Topic Modeling because the text-mining method relies on the assumption known as “bag-of-words”, requiring only words in the lots of texts.

<Figure 7> Example of a Message data (Translated version through Google Translate service)

| |
|---|
| "4-year career, I slip now Android. There are a number of development experience. The one who runs a company based outsourcing and self-service. My portfolio is representative Start-up to a project, who served in the former company. But now it renewed a music streaming service apps. Music Streaming Service https://play.google.com/store/apps/developer?id=IT-TE The apps I did service in the start-up founded. John was now dismantled. Outsource projects https://play.google.com/store/apps/details?id=kr.mothert.word << memorize English words Service, lost App https://play.google.com/store/apps/details?id=kr.tofor.fashionbridge <<. I hope this good ties ^^ Thank you. " |

After translation, I applied topic modeling by *Mallet* and discovered that some words could be labeled as “experience-quality”. Those were “similar”, “project”, “experience”, “quality”, and “career”. Using these, I could make “experience-quality” dummy variable, adding this to bid characteristics. Comparing the results between controlling the topic message variable and not controlling, I could find out that price became more plausible only when it is accompanied with the message variable.

<Table 2 > Results of Topic Modeling with a chunk of Message Data

| topic | parameter | keyword | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0.0131 | design | web | wordpress | theme | work | http | team | university | site |
| 1 | 0.00755 | company | design | development | web | business | system | technology | planning | develop |
| 2 | 0.00496 | development | asp | ms | role | sql | system | technology | main | net |
| 3 | 0.02838 | development | server | web | **experience** | based | company | services | db | design |
| 4 | 0.09035 | development | cost | period | meeting | days | time | project | design | work |
| 5 | 0.00547 | development | **experience** | design | services | variety | electronic | **quality** | web | soft |
| 6 | 0.00904 | development | system | management | web | design | company | **experience** | planning | app |
| 7 | 0.00283 | services | development | planning | web | design | program | app | based | interlocking |
| 8 | 0.00338 | creative | communication | studio | professional | editing | oriented | web | energy | graphics |
| 9 | 0.06906 | project | work | part | proceed | time | planning | projects | related | **experience** |
| 10 | 0.05291 | **experience** | web | years | development | **quality** | team | good | design | results |

Final outcome of topic modeling suggested the possibility that topics other than “experience-quality” domain could be interpreted as other index. Some of them can be classified into three subgroups : “willingness”, “price information”, “period information”. As I mentioned right before, the labeling of each index seems pretty arbitrary. Problem would get severe if topic modeling did not capture the actual perception of buyers to different messages. Besides naïve topic modeling I used, lots of advanced technique has been developed such as local-LDA (Brody et al 2010). Bao et al (2014) utilized the advanced technique of LDA to analyze how risk disclosure would affect the perceived risk of investors.

Using lots of current estimation methods, researchers will be able to make more elaborate index for message data. This is crucial because only thing researcher can control in the quality dimension is message data, which means accurate index for

buyers' perception can enhance the performance of the whole model. Future research can be made with this direction.

# 3. Structural Model

To make use of dynamic programming, researchers would think two parts of model specification : One is for flow utility and the other is for transition probability. Flow utility setup describes how much decision makers would get for that specific behavior. Economics postulates that everyone would pursue the larger utility. Furthermore, in the dynamic situation, rational agents would be expected to maximize their sum of discounted expected utility over the decision period. At this point, the second part, transition probability, is needed to narrow down the possible expectation. With additional economic assumptions, researchers would transform the agents' original problem into more manageable form like Bellman equation. Many approach has been suggested to solve this multi-period controlled stochastic process. In this paper, I followed the CCP approach with EM-algorithm (Arcidiacono and Miller 2011). To begin with, understanding the basic idea in this newly developed method would be helpful for understanding my estimation.

## 3.1 CCP approach with EM-algorithm

### 3.1.1. CCP approach

In general framework of discrete choice models, researchers would set the utility function and estimate the parameters with unobserved variable to allow the discrepancy between model and data. By assigning the distribution to this unobserved variable, researchers can make up the likelihood under the given parameters. Equation (1) describes this static model estimation.

$$
\begin{aligned}
P_{ni} &= \Pr ob(U_{ni} + \epsilon_{ni} > U_{nj} + \epsilon_{nj} \ \ \forall j \neq i) \quad n: agent \ \& \ i, j: choice \\
&= \Pr ob(\epsilon_{nj} < \epsilon_{ni} + U_{ni} - U_{nj} \ \ \forall j \neq i)
\end{aligned} \tag{1}
$$

To compare the dynamic model with the static case, I would jump into the Equation (2). In the dynamic context, researchers would face the choice-specific value function instead of utility function. This is made up with choice-specific utility function at time t and the maximum value of optimized value function given that choice at time t. And to deal with the discrepancy between parameter-mapping space and the real data, unobserved variable is embedded in the model like noise term. Like the above case, researcher would set the likelihood and estimate the most plausible parameters.

$$
\begin{aligned}
P_{ni} &= \Pr ob(v_{ni} + \epsilon_{ni} > v_{nj} + \epsilon_{nj} \ \ \forall j \neq i) \quad where \ v_j(x) = u_j(x) + \beta \int w(x') f(x' \mid x, a = j) dx' \\
&= \Pr ob(\epsilon_{nj} < \epsilon_{ni} + v_{ni} - v_{nj} \ \ \forall j \neq i)
\end{aligned} \tag{2}
$$

It looks quite similar to the static problem, but dynamic problem would require deducting the value function. One of the ways is to set the ex-ante value function like $w(x) = \int V(x,\epsilon) g(\epsilon) d\epsilon$, change the value function problem into x-space problem, and finally derive the ex-ante value function using the fixed point algorithm (Rust 1988). So finding the vector-valued function $w(x)$ is essential part for solving the dynamic programming. Value function iteration, or successive approximation, needs lots of computation time for each different value of parameter. Whenever the parameter changes to find the maximum likelihood function, computer should find value function by fixed point algorithm. To make a detour and reduce the computational burden, Hotz and Miller(1993) developed the Conditional Choice Probability method (CCP). Finding the function $w(x)$ is core part. In the framework of CCP, researchers directly estimate this function from the data.

$$
\begin{aligned}
w(x) = E_\epsilon[V(x,\epsilon) \mid x] &= \int V(x,\epsilon) G(d\epsilon) \\
&= \log(\exp(v_k(x)) \frac{\Sigma_j \exp(v_j(x)}{\exp(v_k(x))}) + \gamma \\
&= v_k(x) - \log(\Pr(a = k \mid x)) + \gamma
\end{aligned} \tag{3}
$$

Equation (3) shows one of the intuitions behind this CCP approach. Ex-ante value function can be decomposed into two parts. $v_k(x)$ and $\log(\Pr(a = k \mid x))$ are the backbone of ex-ante value function. And this decomposition holds for every possible action "k". When researchers choose action k on which decision maker must escape from the decision problem, like purchasing the product and exiting the market, then $v_k(x)$ becomes just the flow utility. The name of CCP stems from the remaining part $\log(\Pr(a = k \mid x))$. For every possible reachable state x, if researchers can calculate this conditional probability, then they can use it as the substitute for the value function.

But when it comes to the problem that involves the unobserved heterogeneity "z", then this approach does not hold. This is because, simply put, $\log(\Pr(a = k \mid x))$ change into $\log(\Pr(a = k \mid x, z))$, making the direct estimation impossible (Researchers do not have data about z ). So this had remained the dead end until the recent research. To deal with this problem, Arcidiacono and Miller (2011) adapted the EM-algorithm and made a great breakthrough. Before explaining this approach, it would be helpful to examine the brief idea behind the EM-algorithm.

3.1.2 CCP approach with EM-algorithm

To begin with, EM-algorithm was originally developed to handle the missing value problem. There is nothing to say about the missing variable for which researchers do not have any information. If researchers speculate the distribution of missing variable and the relationship between unobserved variable and observed variable, then they would focus on the main problem by integrating out this missing part (Expectation). After that, they can estimate the parameters (Maximization). Estimated parameters are used to elaborate the distribution of missing value. After iteration of these steps, researchers would arrive at the same result from other method like MLE. It is quite useful when evaluating the likelihood or updating the likelihood is very difficult. By splitting the problem and solving the smaller problem, EM-

algorithm helps to find the solution within more feasible time. Equation (4) and Equation (5) briefly describe this procedure.

$$h(z \mid x, \theta) = \frac{P(x \mid z, \theta) f(z \mid \theta)}{P(x \mid \theta)} \tag{4}$$

$$\Xi(\theta \mid \theta^t) = \int h(z \mid x, \theta^t) \log(P(x \mid z, \theta) f(x \mid \theta)) dz \tag{5}$$

Variable x represents the observed data and z means the unobserved data. Using the Bayes theorem, the distribution of z would be equated like (4). Using this distribution as the weighting function, we can interpret the RHS of (5) as the weighted average of likelihood. Then, we can find the parameters to maximize (5) and update the (4) again.

Applying this procedure to CCP with unobserved heterogeneity, researchers can take a long way around the missing value and arrive at the structural parameters. This part would be explained in the estimation part.

## 3.2 Model setup

### 3.2.1. State variables

The first thing to set up in the dynamic programming is to specify the state variable. There are two kinds of state variable : One is observed and the other is unobserved. And each state variable can be divided further when it comes to time-varying or not. In the case of observed and time-constant state variable, I can extract the below variables in the dataset.

<Table 3> Observed and time-constant state variable

| |
|---|
| $PP_i$ : expected project price written by buyer i |
| $PT_i$ : expected project term written by buyer i |
| $EN_i$ : duration between project register date and enrollment date on this company. |
| $Pb_i$ : dummy whether first purchase or not |
| $PI_i$ : chracteristic of project |

This variable covers the characteristics of buyer and the characteristics of project. These variables with unobserved buyers' type will affect the utility of waiting. Next

observed state variables change upon each decision period. All of time-varying state variables are related to bid arrival. As time goes by, different kind of bids would be put in the hand of buyers. And these arrived bids have different characteristics such as bid price, bid period. In addition, buyers would consider the origin of bid, the sellers, which made me include the sellers' characteristic. And the feature of message data delivered by sellers to buyers would be incorporated into this group. The outcome from topic modeling is used at this step.

<Table 4> Observed and time-varying state variable

| |
|---|
| $bp_{ij}$ : *bid price of seller(freelancer) j for project i* |
| $bt_{ij}$ : *bid period of seller(freelancer) j for project i* |
| $ba_{ij}$ : *ability of seller(freelancer) j for project i* |
| $I(m_{ij}=0)$ : *dummy whether seller j uses [topic] message or not* |

Then left things are related to unobserved state variables. These things should be united with researchers' assumption about their distribution. Before explaining this assumption, it should be done first to classify these to two groups based on time-dependence. First one is the unobserved and time-constant state variable. This would be interpreted as the type of decision maker, or buyer. This type would not vary depending on time change. Following the Yoganarasimhan (2013), I also split this type into two parts : $\eta_i$ and $s_i$. $\eta_i$ is the parameter that explains the difference between buyers in the perspective of bid arrival. Because of this variable, different buyers would receive different number of bids for each time. So this parameter has the effect only on transition probability, not on flow utility. In contrast, $s_i$ would be closer to the definition of unobserved heterogeneity. This state variable would influence the behavior of agents, which means this variable should be included in the flow utility function. In other words, this variable has the role to explain the systematically different behavior between the buyers who are totally the same with

respect to the observed state variable. This "type" variable has the responsibility for using the EM-algorithm. Second one is the unobserved and time-varying state variable : $\epsilon_{it}$ . This variable would absorb other source of variation in the data that cannot be captured within the existing model. And I assume that this variable would follow the general extreme value distribution to use the nested logit framework for choice set {wait, cancel, select j}

3.2.2. Flow Utility setup

Researchers have to formulate the utility function for describing the agents' behavior with respect to the state variable. To do this, it is needed to specify the utility function for every possible state and every possible behavior.

$$U(d_{it}, x_{it}, s_i, \epsilon_{it}) = u(d_{it}, x_{it}, s_i) + \epsilon_{it}(d_{it})` \quad where\ d_{it} : choice\ of\ agent\ i\ at\ time\ t$$

When this framework is applied into this online freelance market situation, flow utility would be expressed like the below table.

<Table 5> Flow utility according to the behavior

$$\begin{aligned}
&u(1, x_{it}, s_i) = W_{iw}(x_{it}, s_i) = W_{iw}(x_i) \\
&= \alpha_{w1} + \alpha_{w2} PP_i + \alpha_{w3} PT_i + \alpha_{w4} EN_i + \alpha_{w5} Pb_i + \alpha_{w6} PI_i \\
&u(2, x_{it}, s_i) = 0 \\
&u(bid_j, x_{it}, s_i) = W_{ib}(x_i, s_i) + Y_{ij}(x_{it}) \left[ \begin{array}{l} W_{ib}(x_i) = \alpha_{b1} + \alpha_{b2} PP_i + \alpha_{b3} PT_i + \alpha_{b4} EN_i + \alpha_{b5} Pb_i + \alpha_{b6} PI_i + \alpha_{b7} I(s_i = 1) \\ Y_{ij}(x_{it}) = \beta_1 \ln(bp_{ij}) + \beta_2 \ln(bt_{ij}) + \beta_3 \ln(ba_{ij}) + \beta_4 I(m_{ij} = 0) \end{array} \right.
\end{aligned}$$

In this case, choice "1" means "waiting" and choice "2" means "canceling the project". In the discrete choice framework, researchers could only identify the relative size of utility so should normalize the utility function. I choose the canceling utility as the baseline utility and normalize it as 0. This should be done with canceling option to take advantage of "Finite dependence" property in Arcidiacono and Miller(2011). And the utility for "waiting" is the function of observed and time-constant state variable like the feature of project, the characteristics of buyers. For example, it is possible that the repetitive buyers would have high level of utility from

waiting. These kinds of things are mirrored in the utility function. Choice-related utility is disassembled into two parts. To utilize the nested approach, overall utility for "choice" is $W_{ib}(x)$ and bid-specific utility given "choice" is $Y_{ij}(x)$. One thing special in the overall choice utility is the inclusion of unobserved buyers' type variable $I(s_i = 1)$. Therefore, unobserved heterogeneity represents the tendency to select "choice" options. $Y_{ij}(x)$ governs the behavior that buyers would choose the certain bid among the possible bid sets when he or she determines to choose the bid.

3.2.3 State transition probability

To solve the dynamic programming, in addition to flow utility setup, researchers should specify how the aspects of state would change in terms of buyers' side. This is because buyers would anticipate the process of change in time-varying state variable and make the optimal choice based on this expectation. So researchers should set the most realistic assumption about the transition probability. In this part, I heavily borrow the setup in Yoganarasimhan (2013). But I explicitly set the distribution of message usage unlike Yoganarasimhan (2013) and, to focus on this issue, I do not utilize the nonparametric approach for bid-related state variables. Namely, I approach the problem only with parametric assumptions. First of all, future bid arrival is modeled with Poisson distribution. To capture the difference between buyers in bid arrival and to reflect time-varying arrival, I estimate the Poisson parameters for each project. Equation (6) shows the parametric assumption I used.

$$h_p(b_{it} \mid z_t, \eta_i, \theta_p) = \frac{exp(-\lambda_{it})(\lambda_{it})^{b_{it}}}{b_{it}!} \quad \begin{array}{l} b_{it} : the\ number\ of\ price\ arrival\ at\ time\ t \\ \eta_i : parmeters\ for\ poisson\ process \\ \lambda_t (= z_t\eta_{i1} + z_t^2\eta_{i2} + z_t^3\eta_{i3}) : \text{time dummy}(z_t)\ and\ poisson\ process\ parameter \end{array} \tag{6}$$

Other than bid arrival, there are bid-specific characteristics that directly have an influence on the buyers' utility. Different from target paper, parametric approach is used at this part. This is because my dataset is slightly dissimilar with Yoganarasimhan (2013). In target paper's company, buyers should indicate the

maximum bid level they would bear. But the company that gave me the data allows the buyers to suggest the mean level of price they expected to pay for (project price). And the data was generated around this suggestive value like the Normal distribution. Putting this feature into my model, I assume that bid price and bid period would be expected to occur around the project price and period in the Normal distribution. Equation (7) describes these parametric assumptions.

$$\begin{aligned} &(\tilde{b}p_q) \sim N(\mu, \Sigma_p)\ where\ q \in \{1,..,b_{it}\} && \tilde{b}p_q : bid\ price\ suggested\ by\ q \\ &(\tilde{b}t_q) \sim N(\mu, \Sigma_t)\ where\ q \in \{1,..,b_{it}\} && \tilde{b}t_q : bid\ period\ suggested\ by\ q \end{aligned} \tag{7}$$

And finally, message usage is assumed to follow the logit distribution. Information about the sellers' profile or sellers' characteristics is used at this part as independent variable to explain the topic-specific message usage dummy variable.

## 4. Estimation

To estimate the parameters, I specify the choice-specific value function. Equations (8) is the choice-specific value function for each choice.

$$\begin{aligned} &v(bid_j, x_{it}, \eta_i, s_i) = u(bid_j, x_{it}, s_i) = W_{ib}(x_i, s_i) + Y_{ij}(x_{it}) \\ &v(2, x_{it}, \eta_i, s_i) = u(2, x_{it}, s_i) = 0 \\ &v(1, x_{it}, \eta_i, s_i) = u(1, x_{it}, s_i) + \delta \int V'(x_{it+1}, \eta_i, s_i) f(x_{it+1} \mid 1, x_{it}, \eta_i) dx_{it+1} \end{aligned} \tag{8}$$

One thing to notice is that selecting "choosing bid j" or "canceling the project" extricate the decision maker, buyers, from the decision process. This "Terminal" property is essential for which researchers do not have to worry about choice-specific value function with regard to these choices. It is just the same with flow utility. The only one is the waiting-specific value function that bothers the estimation. To analyze this choice-specific value function in detail, the key part of estimation depends on $V'(x_{it+1}, \eta_i, s_i)$. If we define that value function, then likelihood would be specified like Equations

(9) with the help of extreme value distribution.

$$
\begin{aligned}
P(1 \mid x_{it}, \eta_i, s_i) &= \frac{exp[v(1, x_{it}, \eta_i, s_i)]}{1 + exp[v(1, x_{it}, \eta_i, s_i)] + \exp[W_{ib} + \sigma I(x_{it})]} \\
P(2 \mid x_{it}, \eta_i, s_i) &= \frac{1}{1 + exp[v(1, x_{it}, \eta_i, s_i)] + \exp[W_{ib} + \sigma I(x_{it})]} \\
P(bid_j \mid x_{it}, \eta_i, s_i) &= \frac{\exp[W_{ib} + \sigma I(x_{it})]}{1 + exp[v(1, x_{it}, \eta_i, s_i)] + \exp[W_{ib} + \sigma I(x_{it})]} \frac{\exp(Y_{ij} / \sigma)}{\sum_k \exp(Y_{ik} / \sigma)} \\
& where \;\; I(x_{it}) = \ln(\sum \exp(Y_{ij} / \sigma))
\end{aligned}
\tag{9}
$$

So the problem converges to the estimation of $V'(x_{it+1}, \eta_i, s_i)$. By setting the baseline utility setup as "canceling", I can utilize CCP formula like this:

$$V'(x_{it}, \eta_i, s_i) = \gamma + v(2, x_{it}, s_i) - \ln(p(2 \mid x_{it}, \eta_i, s_i)) = \gamma - \ln(p(2 \mid x_{it}, \eta_i, s_i))$$

Relying on the assumption that $u(2, x_{it}, s_i) = 0$ and finite dependence property, I can claim that $v(2, x_{it}, s_i) = 0$. Only thing left in the estimation of value function is canceling probability for every state. EM-algorithm was used at this part because I don't have data about $s_i$.

4.1. Estimation of Transition parameter

Equation (10) represents the full log likelihood function to contain the components in Equation (9). And $\eta_i$ governs the transition probability and has no direct influence on utility. This makes it possible to split the likelihood into two parts like Equation (11). So I can estimate the transition probability ( $\ln\left[\prod_t f(x_{it} \mid d_{it-1}, x_{it-1}, \eta_i)\right]$ ) parameters first, followed by structural estimation. This is done in relatively easy ways with Maximum likelihood estimation.

$$L_i(\theta \mid d_i, x_i) = \ln\left[\sum_{k=1}^{S} \pi_k \left[\prod_t (\Pr(d_{it} \mid x_{it}, \eta_i, s_i = k)^{I(d_{it})} f(x_{it} \mid d_{it-1}, x_{it-1}, \eta_i))\right]\right] \tag{10}$$

$$L_i(\theta \mid d_i, x_i) = \ln\left[\sum_{k=1}^{S} \pi_k \left[\prod_t (\Pr(d_{it} \mid x_{it}, s_i = k)^{I(d_{it})})\right]\right] + \ln\left[\prod_t f(x_{it} \mid d_{it-1}, x_{it-1}, \eta_i)\right] \tag{11}$$

4.2. Estimation of Structural parameter

For given parameters estimated at 5.1 ($\hat{\eta}_i, \Sigma$ (Estimated transition parameter)), I carry out the estimation procedure, which is explained the tables below.

<Table 6> Simulation Procedure : step 1

| |
|---|
| Set the initial value which is needed in the next steps. (Discount rate is fixed at 0.9)<br>$\pi^1 = \{\pi^1{}_1, \pi^1{}_2\}$ : initial value of ratio of each type in the sample<br>$\{\theta^1 : \alpha^1, \beta^1, \sigma_1\}$ : initial parameter value |

<Table 7> Simulation Procedure : step 2

For each decision makers, using the Bayes theorem provides the individual type probability.

$$q_i^2(k \mid d_i, x_i; \theta^1, \pi^1) = \frac{\pi_k{}^1 \prod_{t=1}^{T_i} [\Pr(d_{it} \mid x_{it}, \hat{\eta}_i, s_i = k, \theta^1)^{I(d_{it})}]}{\sum_k \pi_k{}^1 \prod_{t=1}^{T_i} [\Pr(d_{it} \mid x_{it}, \hat{\eta}_i, s_i = k, \theta^1)^{I(d_{it})}]} \quad \forall k$$

In this case, k means the type of unobserved heterogeneity that change the flow utility of option “choosing bid”. In other words, researchers would not know whether buyers i belongs to type1 (s=0) or type2 (s=1). Thanks to Bayes theorem, it is possible to express this probability like the above formula. To aggregate this information gives the overall ratio of type, which would be used at next iteration.

$$\pi^2{}_k = \frac{\Sigma_i q_i^2(k \mid d_i, x_i; \theta^1, \pi^1)}{N} \quad \forall k$$

<Table 8> Simulation Procedure : step 3

Using the estimated individual type ratio, q, I can make a conjecture of conditional choice probability.

$$p^2(2 \mid x, s) = \frac{\sum_n \sum_t d_{2nt} q_{ns}^{(m+1)} I(x_{nt} = x)}{\sum_n \sum_t q_{ns}^{(m+1)} I(x_{nt} = x)}$$

This weighted formula with the results of step2 is the estimated conditional canceling probability at this step. This is possible due to the information on unobserved state variable s at the step2. If the sample size is small, this bin estimator could be replaced with weighted logit formula.

<Table 9> Simulation Procedure : step 4

Now $V'(x_{it+1}, \eta_i, s_i)$ is derived with the form of $p^2(2 \mid x_{it}, \hat{\eta}_i, s_i = k, \rho_i^2)$. This means that given parameters, I can evaluate the value of

$$v(1, x_{it}, \eta_i, s_i) = u(1, x_{it}, s_i) + \delta \int V'(x_{it+1}, \eta_i, s_i) f(x_{it+1} \mid 1, x_{it}, \eta_i) dx_{it+1} \ .$$

One thing that obstruct the evaluation is the existence of integrals. After solving this integral, I can get the expected value function. To deal with this problem, one of way to proceed is using the simulation in the context of Monte Carlo method. Before the structural estimation, there are estimated parameters to govern the transition process, which helps me to generate the simulation data. With this simulated data, I can solve the integral parts, evaluating the expected value function and waiting-specific value function.

<Table 10> Simulation Procedure : step 5

$$P(1 \mid x_{it}, \eta_i, s_i) = \frac{exp[v(1, x_{it}, \eta_i, s_i)]}{1 + exp[v(1, x_{it}, \eta_i, s_i)] + \exp[W_{ib} + \sigma I(x_{it})]}$$

$$P(2 \mid x_{it}, \eta_i, s_i) = \frac{1}{1 + exp[v(1, x_{it}, \eta_i, s_i)] + \exp[W_{ib} + \sigma I(x_{it})]}$$

$$P(bid_j \mid x_{it}, \eta_i, s_i) = \frac{\exp[W_{ib} + \sigma I(x_{it})]}{1 + exp[v(1, x_{it}, \eta_i, s_i)] + \exp[W_{ib} + \sigma I(x_{it})]} \frac{\exp(Y_{ij} / \sigma)}{\sum_k \exp(Y_{ik} / \sigma)}$$

With these probabilities and individual type probability, I can reach the final step of EM-algorithm

$$\theta^2 = \arg\max_\theta \sum_i \sum_t \sum_s q^2(k \mid d_i, x_i; \theta^1, \pi^1) \ln[\tilde{p}^2(d_{it} \mid x_{it}, \hat{\eta}_i, s_i = k, \theta)^{I(d_{it})}]$$

<Table 11> Simulation Procedure : step 6

Now we have new updated parameter $\theta^2$. Then, the updated parameter is going to replace the $\theta^1$ at step2, making other variables like q and pi newly estimated. So the procedure from step2 to step 5 will be repeated until the change hits the tolerance level. The final value would be the same with other estimation procedure like MLE.

# 5. Results

| | Significant | | insignificant |
|---|---|---|---|

<Table 12> Estimated parameters for each models 1~4

| | Model1 | SE | Model2 | SE | Model3 | SE | Model4 | SE |
|---|---|---|---|---|---|---|---|---|
| w_con | 1.9027 | 0.68407 | 5.76289 | 1.38826 | 5.7541 | 1.38833 | 5.7595 | 1.38832 |
| w_log(pp) | -0.298 | 0.11183 | -0.38715 | 0.11587 | -0.38639 | 0.11587 | -0.38677 | 0.11587 |
| w_log(pt) | 0.1460 | 0.1410 | 0.28120 | 0.14536 | 0.27989 | 0.14540 | 0.28003 | 0.14540 |
| w_fp | 0.0359 | 0.18916 | -0.03358 | 0.1849 | -0.03219 | 0.18498 | -0.03292 | 0.18497 |
| w_en | 0.0131 | 0.12439 | 0.10258 | 0.11988 | 0.10239 | 0.11989 | 0.10341 | 0.11988 |
| w_cha | 0.5184 | 0.15002 | 0.50177 | 0.14887 | 0.50202 | 0.14888 | 0.50121 | 0.14892 |
| c_con | -0.7345 | 1.03154 | 6.72290 | 2.25433 | 7.21158 | 2.25550 | 6.97573 | 2.25605 |
| c_log(pp) | -0.2915 | 0.27084 | -0.89043 | 0.19619 | -0.94007 | 0.19666 | -0.91832 | 0.19670 |
| c_log(pt) | 0.9053 | 0.33055 | 0.50256 | 0.24175 | 0.55301 | 0.24185 | 0.53317 | 0.24198 |
| c_fp | 0.9550 | 0.27494 | 0.82242 | 0.27861 | 0.80687 | 0.27905 | 0.81581 | 0.27908 |
| c_en | -0.377 | 0.18447 | -0.01906 | 0.18481 | -0.0206 | 0.18520 | -0.02538 | 0.18516 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| c_cha | 0.5489 | 0.23645 | 1.08173 | 0.25188 | 1.09058 | 0.25219 | 1.07900 | 0.25203 |
| type | 4.6759 | 0.20663 | 6.10026 | 0.19773 | 6.10134 | 0.19863 | 6.09954 | 0.19817 |
| bid_p | -0.295 | 0.22877 | -0.31117 | 0.23049 | -0.32204 | 0.23197 | -0.31514 | 0.23148 |
| bid_t | -0.593 | 0.28122 | -0.57754 | 0.28764 | -0.54059 | 0.28815 | -0.55881 | 0.28881 |
| bid_score | -0.374 | 0.25502 | -0.41701 | 0.25795 | -0.44191 | 0.25807 | -0.42478 | 0.25771 |
| exp_usage | | | | | 0.08813 | 0.03821 | | |
| log(exp_usage) | | | | | | | 0.15148 | 0.11949 |
| cost_usage | | | | | | | | |
| term_usage | | | | | | | | |
| will_usage | | | | | | | | |

<Table 13> Estimated parameters for each models 5~8

| | Model5 | SE | Model6 | SE | Model7 | SE | Model8 | SE |
|---|---|---|---|---|---|---|---|---|
| w_con | 5.765578 | 1.388237 | 5.761678 | 1.388338 | 5.761563 | 1.388247 | 5.754003 | 1.388258 |
| w_log(pp) | -0.38735 | 0.115865 | -0.38691 | 0.115878 | -0.38703 | 0.115868 | -0.38636 | 0.115866 |
| w_log(pt) | 0.280455 | 0.14538 | 0.280837 | 0.145373 | 0.280689 | 0.145373 | 0.280308 | 0.14538 |
| w_fp | -0.03344 | 0.184947 | -0.03391 | 0.184888 | -0.03272 | 0.184956 | -0.03354 | 0.184881 |
| w_en | 0.103887 | 0.119882 | 0.102652 | 0.119882 | 0.102893 | 0.119889 | 0.102866 | 0.119872 |
| w_cha | 0.501734 | 0.148915 | 0.501189 | 0.148892 | 0.50153 | 0.148893 | 0.501227 | 0.14889 |
| c_con | 6.734594 | 2.255214 | 6.711409 | 2.254331 | 6.733647 | 2.256123 | 7.32383 | 2.25748 |
| c_log(pp) | -0.88769 | 0.196323 | -0.88917 | 0.1962 | -0.8916 | 0.196377 | -0.94511 | 0.196993 |
| c_log(pt) | 0.503729 | 0.241995 | 0.501326 | 0.241747 | 0.503155 | 0.241901 | 0.56873 | 0.242288 |
| c_fp | 0.814346 | 0.279042 | 0.822001 | 0.278657 | 0.818943 | 0.278703 | 0.788745 | 0.279671 |
| c_en | -0.01753 | 0.185021 | -0.01862 | 0.184849 | -0.02037 | 0.184837 | -0.03681 | 0.185332 |
| c_cha | 1.067044 | 0.251714 | 1.080068 | 0.251887 | 1.086506 | 0.252138 | 1.079402 | 0.252164 |
| type | 6.124141 | 0.198213 | 6.099388 | 0.19772 | 6.092718 | 0.197517 | 6.09383 | 0.198558 |
| bid_p | -0.32485 | 0.231548 | -0.31089 | 0.230514 | -0.31034 | 0.230128 | -0.34091 | 0.213565 |
| bid_t | -0.55936 | 0.288935 | -0.57766 | 0.28772 | -0.58163 | 0.287692 | -0.51371 | 0.290321 |
| bid_score | -0.4011 | 0.258213 | -0.41592 | 0.25817 | -0.4316 | 0.259808 | -0.43534 | 0.259763 |
| exp_usage | | | | | | | 0.101523 | 0.039384 |
| log(exp_usage) | | | | | | | | |
| cost_usage | -0.19194 | 0.169806 | | | | | -0.27211 | 0.185281 |
| term_usage | | | -0.01028 | 0.142038 | | | 0.022535 | 0.156367 |
| will_usage | | | | | 0.115051 | 0.237021 | 0.086059 | 0.241057 |

Table 12 and 13 shows the estimation results for 8 models. Model 1 represents the basic model that combines all the structured data. In Model 2, I transformed the bid price into relative value based on project price. This improves the fitness of model significantly. Variables with “w-” mean the variables in the “wait” nest and variables with “c-” refer to ones in the “choose” nest. The result shows that the project price has an influence on buyer’s utility to choose “wait”. When project size gets bigger, the utility from waiting diminishes. Project characteristics about whether the project belongs to “development” category or “design” category is also significant. Compared to buyers of the project in “development”, buyers posting the design-related projects seem to get more utility from waiting. For “choose” nest, it is a remarkable result that repeated buyers, or customers who use the service of the online freelance company more than one time, show higher utility level from “choose” behavior. Other than that, buyers with bigger project price shows a reluctance in choosing behavior. From Model 3 to Model 8, I examined the significance of each topic in different messages. Only the message dummy indicating the experience-quality usage is significant as you can see in Model 3 and Model 8. Other message dummies like price, period, or willingness dimension are estimated not to be significant. And with experience dummies, price coefficient slightly changes to the direction to be significant. Unfortunately, totally controlling the quality part that buyers would guess seems impossible. With the help of topic modeling, I can extract the quality-related parts in the message dataset. But the topic modeling does not capture all the details in the buyers’ various perception on different messages. Other advanced technique which evolved from Topic Modeling can be applied to generate more accurate topic dummy variables.

# 6 Conclusion

Following a dynamic structural estimation procedure, which was developed by Arcidiacono and Miller (2011) and applied by Yoganarasimhan (2013), I analyze the characteristics of initial online freelance market. This market suffered from information asymmetry problem due to the lack of supplementary reviewing process. To alleviate the problem, the Korean company introduced a messaging system for freelancers to have a chance to appeal themselves. But the effect of this remedy has not been measured until this study. To include the unstructured data like message data in CCP estimation, I use the Topic Modeling method and insert the feature of message into the framework of dynamic programming. In this way, I can quantify the value of message as well as other bid-related features such as price. I discover that buyers could be insensitive to price especially in the initial online market which is full of uncertainty. This might be because quality information was mingled with price. After controlling the message data that contained the quality information, I can observe a more plausible price coefficient.

<Reference>

Arcidiacono P, Ellickson PB (2011) Practical methods for estimation of dynamic discrete choice models. Annual Rev. Econom.3(September):363–394.

Arcidiacono P, Miller RA (2011) Conditional choice probability estimation of dynamic discrete choice models with unobserved heterogeneity. Econometrica 79(6) 1823–1867.

Bajari, Parick, Denis Nekipelov, Stephen P. Ryan, and Miaoyu Yang. 2015. “Demand Estimation with Machine Learning and Model Combination.” NBER Working Paper.

Bao, Y., Datta, A., 2014. Simultaneously discovering and quantifying risk types from textual risk disclosures. Manage. Sci. 60 (6), 1371–1391.

Brody S, Elhadad N (2010) An unsupervised aspect-sentiment model for online reviews. The 11th Annual Conf. North Amer. Chapter of the Assoc. Comput. Linguistics (ACL, Stroudsburg, PA), 804–812.

Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. J. Machine Learn. Res. 3:993–1022.

Chung DF, Steenburgh T, Sudhir K (2013) Do bonuses enhance sales productivity? A dynamic structural analysis of bonus-based compensation plans. Marketing Science, 2013

Nevo A, Whinston MD (2010) Taking the dogma out of econometrics: Structural modeling and credible inference. The Journal of Economic Perspectives, 2010

Newey WK, McFadden D (1994) Large sample estimation and hypothesis testing. Handbook of econometrics, 1994

Rust J (1987) Optimal replacement of GMC bus engines: An empirical model of Harold Zurcher. Econometrica 55(5):999–1033.

Rust J (1988) "Maximum Likelihood Estimation of Discrete Control Processes," SIAM Journal on Control and Optimization.

Rust J (1994) Structural estimation of Markov decision processes. Engle RE, McFadden D, eds. Handbook of Econometrics, Vol. 4 (North-Holland, Amsterdam), 3082–3139.

Train KE (2008) EM algorithms for nonparametric estimation of mixing distributions. Journal of Choice Modelling, 2008

Bao, Y. & Datta, A. (2014). Simultaneously discovering and quantifying risk types from textual risk disclosures. Management Science, 60 (6).

Yoganarasimhan H. (2013) The value of reputation in an online freelance marketplace, Marketing science Vol. 32, No. 6, November–December 2013, pp. 860–891

Yoganarasimhan, H. (2015). Estimation of beauty contest auctions. Marketing Science, Forthcoming.

# 국 문 초 록

## 초기 온라인 프리랜서 시장 수요 분석 및 메시지 시스템의 효과 분석

본 연구에서는 초기 온라인 프리랜서 시장에서 회사의 주요 고객인 소비자의 수요함 수를 동적 최적화 모형을 통해서 분석하였다. 소비자(buyer)는 프리랜서(seller)의 특징을 살펴보면서 매 시점마다 기다리거나(wait), 취소하거나(cancel), 선택(select)을 하게 된다. 이를 optimal stopping problem으로 구조화를 한 후에 동적 최적화 모형을 적용하였다. 시장의 특성인 정보 비대칭의 문제를 해결하고자 기업들은 다양한 노력을 하였는데, 기존 연구에서는 그 중에서 평가 데이터에 초점을 맞춰서 그 효과를 분석하였다. 하지만 본 연구에 활용된 자료에서는 이러한 평가 데이터가 부재한 초기 시장에서 기업이 대처한 방법의 효과를 분석하였다. 초기 시장에서 기업은 빠르게 정보 비대칭을 완화시킬 방법이 필요했고 이를 위해서 메시지 시스템을 도입하였다. 이 메시지를 통해서 프리랜서는 소비자에게 자신의 경험이나 의지를 전달할 수 있고 이를 통해 정보 비대칭을 완화할 수 있으리라는 것이다. 하지만 본 연구 이전까지 이 메시지의 효과를 추정한 연구는 없었다. 이는 메시지라는 것이 비정형 데이터이기 때문인데, 이를 해결하고자 본 연구에서는 토픽 모델링을 통해서 비정형 데이터를 변환시킨 후 기존의 데이터와 결합하여 동적 최적화 모형을 통해 결과를 도출하였다. 그 결과 소비자는 이러한 텍스트 데이터 중에서도 경험과 관련된 점에 대해서만 유의미하게 반응함을 알 수 있었다. 그리고 이러한 텍스트 데이터를 분석에 포함시키지 않을 경우 소비자가 마치 가격에 민감하게 반응하지 않는 것처럼 잘못 추정 될 수도 있음을 보였다.

**주요어** : 구조적 접근; 동적 최적화; 텍스트 데이터; 토픽 모델링; 온라인 프리랜서 시장
**학 번** : 2014-20435