第一页为封面页

参赛队员姓名：杜晨牧

中学：BNDS（Beijing National Day School）

省份：北京市

国家/地区：中国

指导教师姓名：王勇、杨菲

指导教师单位：清华大学经济学研究所、北京市十一学校国际部

论文题目：Stable Matching Algorithms of the Two-sided Data Trading Market

Stable Matching Algorithms of the Two-sided Data Trading Market

Chenmu Du

Abstract

With the advent of the digital economy, the significance of data as a production component has become increasingly prominent. It is well noted that the data element in producing value is not just the data alone; data can only produce value when combined with algorithms and models to produce "information", such as description, diagnosis, prediction, and decision. Based on the ascending relationship between data and information, we aim to study how information is generated through the matching process between data elements. Specifically, we take the data matching mechanism of the big data trading platform as the research object and study how the big data trading platform, as an intermediary, could match the behavioral data (or performance data) of the demand side and the feature data (or background data) of the supply side to generate information needed by the data demanders. In this study, we present a matching framework for studying the data trading market from a market viewpoint that explicitly combines essential data trading features: decision-making from both sides. By categorizing data into four categories based on its type: behavioral data, performance data, feature data, and background data, we build the one-to-one and one-to-many two-sided matching models in the data trading market, and establish the stable matching algorithms accordingly. Originated from the two-sided matching theory, we believe these algorithms provide a novel approach to data matching problems, which could be further extended to other fields of the data trading market as well.

1 Introduction

Two-sided matching problems, first introduced by Gale and Shaply in 1962, and further developed by Alvin Roth, have been widely researched in economics. The buyers and sellers in two-sided matching markets have preferences about whom they deal with on the other side of the market. The firms, for example, that compete for college graduates are concerned about whom they hire. Meanwhile, college graduates cannot simply demand the firm they choose, as the firm must also choose him as well. Such a two-sided market clears not only through payment.

Data has recently been perceived as a new oil or currency in the digital world in the big data era. It has been identified as a new form of production element. The big data market has also been identified as one of the five core production markets in China. However, with the increasing diversity of data types and volumes, data trading platforms are facing a number of difficulties. While numerous challenges are associated with making such a trading market operational, one major problem is: How to match data demanders with certain information needs and data suppliers with data privacy concerns?

1.1 Overview

1.1.1 Data and Information

The DIKW model proposed by Ackoff (1989) provides a suitable analytical framework for understanding data as a fundamental but complex concept in information science, which is inseparable from the conceptual analysis of information, knowledge, and wisdom, and suggests that data, information, knowledge, and wisdom have an ascending relationship. That is, data is derived from raw observation and measurement, information is derived from data, knowledge is summarized from information, and wisdom is sublimated via dialogue and self-reflection among wise people. As a result, information, knowledge, and wisdom are "higher order" concepts compared to data. Furthermore, the worth of data, information, knowledge, and wisdom differ at various levels. The value of data is expressed at the micro level in the enhancement of users' utility, while at the macro level in the information, knowledge, and wisdom retrieved from data, which can have a multiplier effect and assist in improving total factor productivity. As a result, we study how data elements generate information through the matching process based on the ascending relationship between data and information.

1.1.2 Big Data Trading Market

As the amount of data expands quickly, enormous databases with extensive content and depth become more common. The basic objectives of big data trading fall into two groups. On one hand, the data trading process should maximize the earnings of data suppliers. On the other hand, this method must suit the informational needs of demanders. Undoubtedly, this is a win-win scenario for both suppliers and demanders.

Big data is the foundation for the productivity resolutions of the future generation of data owners. Through the services they provide, data owners like Tencent and Alibaba amass enormous volumes of data. Clearly, developments in big data analytics supplemented by machine learning and data mining approaches provide huge value for these organizations. However, not all firms are able to gather the requisite data, since the collection of big and exhaustive datasets requires substantial infrastructure investment and sustained work. Data owners have a strong desire to trade their own datasets with others in

order to provide services, boost productivity, and maximize the value of data. In the meantime, firms need information to discover new business opportunities, consumer values, and customers in environments of high competition. Therefore, data demanders also have a strong urge to acquire data for reasons of information gatherings.

Privacy is a crucial consideration for both suppliers and demanders. To safeguard demanders' privacy, certain personal information should be masked throughout the data trading process. Similarly, privacy is plainly vital for data suppliers as well. For privacy protection, people often use both legal oversight and technical measures, such as copyright laws, encryption licensing, and so on, among which cryptography-based approaches are prevalent. For example, Fan Liang et al.(2018) propose a concealing design strategy to encrypt and conceal a part of the data from the original source. This offers a basic foundation for the data trading market in general.

1.1.3 Two-sided Matching Theory

Gale and Shapley (1962) introduced the two-sided matching problem in "College Admissions and the Stability of Marriage" in 1962. This work investigates the notion of marriage matching, the idea of stable marriage matching and its existence, the Deferred Acceptance Algorithm, and Pareto optimality, and proposes the concept of two-sided matching. In 1985, Alvin Roth (1985) proposed the notion of two-sided matching, which relates to how to match topics in two distinct finite sets with the objective of matching each subject to a suitable subject on the other side. Shapley and Roth were awarded the Nobel Prize in Economics in 2012 for their contributions to the solution of the two-sided matching issue.

There are a few characteristics of two-sided matching: (1) two-sided matching is the pairing of two subjects from a finite set. (2) two-sided matching can take place with or without the use of an intermediary. (3) Each party has its own set of requirements, and the matching result must meet those requirements. (4) According to the number of matches between the two parties involved in the market activity, a two-sided matching theory is divided into three categories. When both parties choose one matching object, it is a one-to-one two-sided matching theory; when one party chooses more than one matching object, it is a one-to-many two-sided matching theory; when both parties choose more than one matching object, it is a many-to-many two-sided matching theory.

1.2 Related Works

Most research on big data trading focuses on data trading mechanisms, data trading development paths, data trading property rights, and data trading legal challenges. Fan Liang et al.(2018) examined the current state of big data research from different perspectives; Luo Pinliang et al.(2014) investigated the necessity and strategy of bilateral pricing based on merchants and consumers in the context of online shopping platforms in China; Xiong and Tang (2021)

outlines data asset features such as replicability and value certainty, as well as data circulation, trading, and pricing of digital items and data products. Rather of forcing organizations to actively exchange data, a variety of data trading systems have been created; Arnold and others (2007) considered the information characteristics of buyers and sellers, risk valuation, and price discrimination and established a search model to investigate the participation of big data buyers in the transaction price setting process under the conditions of an incomplete information market. Ming, Xu, and Wang et al.(2015) developed a systematic framework for trading data traffic in mobile terminals; Wang and Yang (2007) studied the two-sided matching problem in the process of outsourcing information systems for knowledge trading and considered six attributes that both parties could evaluate. Chen Xi and Fan Zhiping (2012) investigated the two-sided matching choice problem in knowledge trading to get matching satisfaction. They suggested a multi-attribute decision model based on the linguistic Choquet integral operator.

2 Model Framework

2.1 Classification of Data Based on Data Types

We broadly classify data into four categories according to data types: behavioral data, performance data, feature data, and background data. Behavioral data include user-driven behaviors such as consumer purchase behavior, web browsing behavior, etc.; performance data refer to the effectiveness data generated by the work behavior and manner within a certain period of time; feature data refer to various types of biometric data such as gender, age, occupation, city, hobbies, etc.; background data include other data which is highly relevant to the required information, such as time, weather, seasonal and other environmental data, etc..

Behavioral data (or performance data) usually matches feature data to form information; when there is background data, background data usually matches behavioral data (or performance data) together with feature data to form information. In our two-sided matching model, behavioral data is provided by data demanders, feature data is extracted from the encrypted database of the big data trading platform, which is provided by data suppliers. Background data can be provided by the data demander or filtered and extracted from the encrypted database provided by data suppliers.

2.2 Two-sided Matching in Data Trading Market

In the study of the two-sided matching problems in the data trading market, the big data trading platform is introduced as an intermediary to match the behavioral data (or performance data) from the demand side and the feature data from the supply side, according to the highest bid that data demanders offer in exchange for certain information, and the lowest price set by the data suppliers. One set of feature data is only allowed to be matched with one behavioral

data from the demand side in the same matching period. The mathematical definition of the two-sided matching problem in data trading market is given below:

Definition 1 (*Two-sided Matching in Data Trading Market*)**.** Let the set of behavioral data provided by the demand side be $Y = \{Y_1, Y_2, \ldots, Y_n\}$, where Y_i denotes the i-th behavioral data, $i = 1, 2, \ldots, n$. Let the set of feature data in the encrypted dataset on the platform be $X = \{X_1, X_2, \ldots, X_m\}$, where X_j denotes the j-th feature data, $j = 1, 2, \ldots, m$. Define the two-sided matching μ as the mapping $\mu : Y \cup X \to Y \cup X$, and $\forall Y_i \in Y, \forall X_j \in X, \mu$ satisfies the following conditions:

1. $\mu(Y_i) \in X \cup \{Y_i\}$, if $\mu(Y_i) = Y_i$, then the behavioral data Y_i is said to have no match.
2. $\mu(X_j) \in Y \cup \{X_j\}$, if $\mu(X_j) = X_j$, then the feature data X_j is said to have no match.
3. if $\mu(Y_i) = X_j$, then $\mu(X_j) = Y_i$.
4. if $\mu(Y_j) = X_i$, then $\mu(X_j) \neq Y_k, \forall k = 1, 2, \ldots, n, k \neq i, Y_k \in Y$.

2.3 One-to-one Matching

To simplify the model, in this section, we first assume that each demander provides behavioral data Y_i that match at most one feature data X_j. The expected bid of the demander is p_i , and the lowest expected price of each feature data X_j set by the data suppliers is q_j, where $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, m$. Note that the feature data can only be selected when $p_i > q_j$. On one hand, the preference for each behavioral data Y_i over the feature data X_j can be expressed as a preference sequence $P(Y_i)$ on the set $X \cup \{Y_i\}$, and is evaluated by the "level of relevance" of the two entries. In other words, for each behavioral data, we rank the feature data by its level of relevance. On the other hand, the preference sequence $P(X_j)$ of each feature data X_j is determined by the expected bid of the demanders.

Definition 2 (*level of relevance*)**.** Whether two types of data are considered correlated depends on two aspects: the level of significance and the correlation coefficient.

(1) Significance level, that is, the p-value. Generally, a p-value less than 0.05 is significant; a p-value less than 0.01 is more significant; and a p-value ¡ 0.001 is of very high level of significance. Specifically, we use $p < 0.05$ as a benchmark in our algorithms.

(2) Correlation coefficient, also known as Pearson Correlation (Pearson correlation coefficient). The correlation coefficient can be a value between -1 and +1. The larger the absolute value of the correlation coefficient, the stronger the relationship between the variables. The Pearson correlation coefficient is calculated as follows.

$$\rho(x,y) = \frac{cov(X,Y)}{\sigma_x \sigma_y} = \frac{X * Y}{|X||Y|}$$

where the covariance and the product of the standard deviations are served as the numerator and the denominator respectively. Since

$$\mu_x = E(X); \sigma_x^2 = E(X - \mu x)^2 = E\left(X^2\right) - E^2(X),$$

the formula of Pearson correlation coefficient can also be written as:

$$\rho(X,Y) = \frac{E(XY) - E(X)E(Y)}{\sqrt{E(X^2) - E^2(Y)}\sqrt{E(Y^2) - E^2(X)}}$$

As a result, we define *level of relevance* of two sets of data as the absolute value of the correlation coefficient of the two, given that $p < 0.05$.

2.3.1 One-to-one Stable Matching

Theorem 1 (One-to-one Stable Matching Algorithm)**.** *For any such one-to-one two-sided data trading market, there is always a stable matching* μ_Y*.*

Proof. The algorithm for generating a stable matching μ_Y in any such data element two-sided trading market is as follows.

Step 1: For each behavioral data of the demand side, first calculate and rank its level of relevance with each feature data of the supply side, and eliminate the feature data with low significance level. The initial matching status of the behavioral data is set to "unpaired."

Step 2: Each "unpaired" behavioral data is first paired with the feature data with the highest level of relevance. Suppose the demander's bid is lower than the lowest expected price of that feature data set by the suppliers, i.e.,$p_i < q_j$, the behavioral data is set to "unpaired" status for this feature data, and vice versa. For the feature data with multiple "pairable" behavioral data, the behavioral data with the highest bid from the demanders is selected to maintain the "paired" status, and the rest of the behavioral data is returned to the "unpaired" status. The rest of the behavioral data is returned to the "unpaired" status. Change the pairing status of all "pairable" behavioral data to "paired" with the feature data they match.

Step 3: For any "unpaired" behavioral data, pair it with the next feature data in the preference sequence, and repeat step 2 until no more behavioral data could be paired. after which the algorithm stops. (Note that the pairing status can be updated if there is a new pairing request for the "paired" feature data, as described in step 2.)

To prove that this is a stable matching, assume that some behavioral data y and some feature data x do not match each other at matching μ_Y, but x has a higher level of relevance to y than the current match. To prove that this is a stable match, we need to show that (x, y) does not constitute a "destruction pair". Since x is more correlated with y than the current match, it must have already been paired with y before x is paired with the current match. Since x and y fail to match under matching μ_Y in the end, (x, y) must not constitute a "destruction pair", and therefore, μ_Y is a stable matching as claimed.

□

Corollary 1.1. *Since behavioral data of the demand side and feature data of the supply side are symmetric in the data trading market, the new algorithm obtained by swapping the behavioral data and feature data in the above algorithm still produces a stable matching μ_X.*

Remark 1. μ_X and μ_Y are not necessarily the same.

Definition 3 (*One-side Optimal Stable Matching*)**.** For a given two-sided data trading market, a matching μ_Y is a *demand-side optimal stable matching* if the relevance of the feature data matched to the behavioral data by matching μ_Y is at least as good as any other match, i.e., $\mu_Y \geq \mu$ for any other stable matching μ. Similarly, matching μ_X is a supply-side optimal stable matching if the demand-side bids matched by matching μ_X are at least as high as any other matching μ'. That is, we have $\mu_X \geq \mu'$ for any other stable matching μ'.

Definition 4 (*Achievable*)**.** For a given two-sided data trading market, the behavioral data Y of the demand side and the feature data X of the supply side are said to be *achievable* for each other if they could be paired with each other in some stable matching.

Corollary 1.2. *The optimal stable matching on the demand side ensures that behavioral data of the demand side is matched with the most relevant available feature data of the supply side, and the optimal stable matching on the supply side ensures that the feature data of the supply side is matched with the most priced available behavioral data of the demand side.*

Theorem 2 (One-side Optimal Stable Matching Theorem)**.** *For a given two-sided data trading market, there will always be a demand-side optimal stable matching μ_Y and a supply-side optimal stable matching μ_X. Further, the matching μ_Y resulting from Theorem 1 is the demand-side optimal stable match. Conversely, the matching μ_X resulting from the new algorithm obtained by swapping the behavioral and feature data in the algorithm in Corollary 1.1 is the supply-side optimal stable matching.*

Theorem 3. *Supply and demand share opposite preferences in the set of stable matching: If μ and μ' are stable matching, then all supply sides will prefer μ at least as much as μ' if and only if all demand sides prefer μ' at least as much as μ, i.e., $\mu > \mu'_Y$ holds when and only if $\mu' > \mu_X$.*

This theorem leads to the following corollary.

Corollary 3.1. *The optimal stable matching on the demand side is the worst for the supply side, i.e., it matches each feature data of the supply side with the lowest preference of all available behavioral data of the demand side; and vice versa.*

2.4 One-to-many Matching

In this section, we assume that each behavioral data Y_i of the demand side can match multiple feature data $\{X_j\}$, i.e., the set of feature data that provide the most information within the affordable prices. Again, the bid offered by the demand side is p_i, and the lowest price of each feature X_j set by the supply side is q_j. The set of features $\{X_j\}$ can be selected when $p_i > \sum q_j$, where $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, m$.

2.4.1 One-to-many Stable Matching

Theorem 4 (One-to-many Stable Matching Algorithm)**.** *For any such two-sided data trading market, there always exists a stable matching μ_Y.*

Proof. The algorithm for generating a stable matching μ_Y in any such two-sided data trading market is as follows:

Step 1: For each behavioral data of the demand side, first calculate and rank its level of relevance with each feature data of the supply side, and eliminate the feature data with low significance level. At the same time, for each feature data, the behavioral data with which it is correlated is sorted in descending order of pricing (eliminating behavioral data with demand-side bids lower than the pricing of the feature data) to form a preference sequence of the feature data.

Step 2: These sequences are put into a sequence processing algorithm consisting of a "matching phase" and an "experimental pairing and updating phase". The first Step (Step 1 : 1) of the matching phase is to see if there are behavioral data and feature data that are each other's top-ranked choices. In this case, for the feature data, the top ranking is the highest priced behavioral data, i.e., the first one in its preference sequence; for the behavioral data Y_i, the top s_i features in its preference sequence is selected. $(s_i = \underset{j=1,2,\ldots,m}{\operatorname{argmax}} \sum q_j \leq p_i)$ as the top-ranking choice. If there are behavioral data and feature data that are each other's top-ranked choices, the matching phase enters into a temporary matching pair. Otherwise, the matching phase enters into Step 2 : 1, where the second position in the feature data preference sequence is compared with the first s_i feature data (the same s_i as above) in the behavior data preference sequence. If no matching is found after any step, the algorithm proceeds to the next step, collectively called Step k : 1, in which the kth feature data in preference sequence is matched with the first s_i features in the behavioral data

preference sequence (the same s_i as above). Once a matching is found after a certain step, the algorithm proceeds to the "experimental pairing and updating phase".

Step 3: when the algorithm moves from the matching phase $k:1$ step to the experimental pairing and updating phase, the $k:1$ matching is experimental, i.e., the behavioral data ranked kth in a certain sequence of feature data preferences is experimentally paired with that behavioral data if it is the same as the top-ranked feature data in that sequence of behavioral data's preferences. The sequence is then updated in the following way: any behavior data that is ranked lower than its experimental match is removed from its sequence (so the sequence of the feature data X_j that is experimentally paired with its kth choice is updated to include only its first k choices), and any behavioral data that is removed from its sequence also removes that feature data from their sequence. (So the updated sequence for each behavioral data includes only those applicants that do not have an experimental match to their more preferred behavioral data.) If a top-ranked feature data is removed from the list of behavioral data, then the slightly lower-ranked feature data goes into the top-ranked range and the top s_i feature data in its preference sequence is reselected ($s_i = \underset{j=1,2,\ldots,m}{\operatorname{argmax}} \; sumq_j \leq p_i$) as its top-ranking choice. When these sequences are updated in this way, the algorithm returns to the "matching phase" and continues to check the updated sequences for new matching. Any new experimental matching created in the "matching phase" replaces the original experimental matching containing the same feature data. It is worth noting that new experimental matching can only improve the experimental matching of the feature data, since all behavioral data ranked further back have been removed. The algorithm ends when no new experimental matching is generated, at which point the trial experimental matching become the final matching.

At the end of the algorithm, each behavioral data Y_i is matched with the top s_i feature data ($s_i = \underset{j=1,2,\ldots,m}{\operatorname{argmax}} \sum q_j \leq p_i$) in the last updated sequence. (This holds because the algorithm does not end when the experimental $k:1$ matching can still be found.) This matching is stable because when any feature data X_j that ranks before the final matching of this behavioral data Y_i is matched in the experimental matching with a behavioral data that ranks before Y_i in the sequence of this feature data, this feature data will be removed from the sequence of Y_i, and thus in the final matching, X_j will be matched with the behavioral data ranked higher than Y_i rather than Y_i. Therefore, this is a stable matching as claimed. □

Corollary 4.1. *Similar to one-to-one two-sided matching, for a given one-to-many two-sided data trading market, there always exists a demand-side optimal stable matching μ_Y and a supply-side optimal stable matching μ_X. Furthermore, the matching generated by algorithm 4 is a demand-side optimal stable matching.*

3 Conclusion

In this paper, we classify data into four categories: behavioral data, performance data, feature data, and background data based on the perspective of information generation. We then develop one-to-one and one-to-many two-sided matching models in data trading market, and design stable matching algorithms accordingly. We believe these algorithms provide a novel approach to data matching problems originating from the two-sided matching theory, which could be further extended to other fields of the data trading market.

For further discussion, we can introduce background data to the two-sided matching model, and construct a one-to-many two-sided matching model with background data, i.e., data that are highly relevant to the behavioral data, but need to be matched together with feature data to form information. In that case, we may have to calculate complex correlation coefficients to define "level of relevance" rather than the Pearson correlation coefficients we used above.

请如实说明：

1. 论文的选题来源、研究背景；

偶然一次机会去北京的大数据交易所，了解了中国的数据和信息相关知识，了解了现在关于数据方面存在许多供需不匹配不平衡的问题。于是和导师讨论了一下是否可以建设一个模型用来匹配供给方和需求方提供的数据。就有了这个选题。

2. 每一个队员在论文撰写中承担的工作以及贡献；

全文在指导老师的辅导下由独立完成。

3. 指导老师与学生的关系，在论文写作过程中所起的作用，及指导是否有偿；

无偿，进行知识点检查补充，使用术语和句式修改。

4. 他人协助完成的研究成果。

无