Stony Brook University

Detecting Fraud in Public Procurement

A Dissertation presented

by

Yajun Wang

to

The Graduate School

in Partial Fulfillment of the

Requirements

for the Degree of

Doctor of Philosophy

in

Economics

Stony Brook University

August 2016

Stony Brook University

The Graduate School

Yajun Wang

We, the dissertation committe for the above candidate for the

Doctor of Philosophy degree, hereby recommend

acceptance of this dissertation

Sandro Brusco - Dissertation Advisor
Professor, Economics Department

Yiyi Zhou - Chairperson of Defense
Assistant Professor, Economics Department

Ting Liu - Committee Member
Assistant Professor, Economics Department

Wei Tan - Outside Member
Vice President, Compass Lexecon

This dissertation is accepted by the Graduate School

Nancy Goroff
Interim Dean of the Graduate School

Abstract of the Dissertation

Detecting Fraud in Public Procurement

by

Yajun Wang

Doctor of Philosophy

in

Economics

Stony Brook University

2016

Fraud in public procurement is a big problem in public sector all over the world, and is very difficult to detect in empirical studies. Common fraud schemes include corruption, collusive bidding, failure to provide the required quality, and false statements, etc. In this paper, I employed game theory, machine learning, and statistical methods to detect fraud risk in Federal Procurement Contract Data, and studied the relationship of fraud, competition and contract types. In the first section, I studied a procurement game and found that if the firms' types are close enough to each other, their strategies regarding whether or not to engage in fraud would tend to be similar. Based on this proposition, in the second section, I implemented One-Class Support Vector Machine method to train the historical data of contractors with fraud records, and developed a classifier. Then I used the classifier to classify and analyze the Federal Procurement Data. In the last section, I applied Logit Regression to the classification outcomes, and the result shows that competition has a small positive relationship with fraud risk. In addition, performance based contracts and flexible-price contracts are more inclined to fraud.

Contents

List of Figures

List of Tables

List of Abbreviations

DOD Department of Defense

DOT Department of Transportation

DUNS Data Universal Numbering System

FAR Federal Acquisition Regulation

FCMD Federal Contractor Misconduct Database

FPDS Federal Procurement Data System

GSA General Services Administration

NE Nash Equilibrium

OECD The Organisation for Economic Co-operation and Development

POGO Project on Government Oversight

SAM System for Award Management

SVM Support Vector Machine

1 Introduction and Literature

Fraud has long been a big problem of public procurement across the world. Common Fraud scheme in procurement includes corruption, collusive bidding, false claim of information, manipulation of bids, etc. All these schemes have the same feature of the contractor failing to provide the quality-price level as required, or as it should be under complete competition.

One example is bribery in public procurement. According to OECD (2007), public procurement accounts for 15 percent of GDP in about 30 countries including Australia, France, Germany, Japan, Mexico, New Zealand, the Slovak Republic, the United Kingdom and the United States. And bribes in transnational business may range from 5 to 25 percent of the value of a contract, or even more. In military supplies the bribe may reach 30 percent in the Gulf region, 10 percent in Africa, 5 to 20 per cent in Latin America and 5 per cent in Taiwan.

In procurement, corruption mostly stems from lack of observability of quality. In most cases, the government cannot evaluate the quality of the goods or monitor the efficiency of the sellers, thus it has to delegate a third-party intermediary to do the job. However, because this third-party agent does not have the mutual objective with the government, and act on its own behalf, corruption sneaks in here. In this case, procurement can be seen as a multi-hierarchy agency problem.

The related literature about corruption and auction can be traced back to Becker and Stigler (1974), Banfield (1975), Rose-Ackerman (1975), Myerson (1981), and Shleifer and Vishny (1993). Tirole (1986) is the first to study the collusion in organizations by constructing a three-tier agency model in which there is one supervisor who monitors the firms for the principal. Firms can collude with the supervisor by providing a bribe, and Tirole showed that if contracts are appropriately made, this kind of collusion could be prevented.

Following Tirole's framework, Kofman and Lawarrée (1993) established a more complicated hierarchical agency model with one or more auditors between the principal and the agent. There are both internal auditors and external auditors, and only internal auditors collude with the agent. Their result shows that the optimal contract will need random external audits. Olsen and Torsvik (1998) studied a dynamic version of the three-tier agency model and showed that for an intertemporal commitment problem, corruption can be beneficial for the principal. Auriol (2006) also used this framework to study corruption in public procurement, and he focused on the different costs between "active bribery" to get a trading advantage and complying with a demand for a bribe in order to avoid being excluded from the trade.

Laffont and Martimort (1997, 1998, 2000) have done a lot of work in collusion in the framework mechanism design analysis. They analyzed the problems in which the agents could enter collusive agreements with each other under asymmetric information (Laffont and Martimort, 1997), or the agents have private information in a public environment with correlated types (Laffont and Martimort, 2000). They have also discussed collusion and delegation game by analysis of centralized organizations and decentralized organizations. In these cases they focused on the mechanism design problem of a collusion-proof contract, and used a third-party mechanism designer to implement a side contract to achieve that. Faure-Grimaud et al. (2003) discussed a hierarchical agency model with a supervisor between the principal and the agents, and analyzed both centralized and decentralized structures, i.e., the case when the principal can contract with both supervisor and agents, or the case when principal can only contract with the supervisor. They obtained the result that the two kinds of organizations could achieve the same outcome. Che and Kim (2006, 2009) have also focused on collusion-proof implementations in mechanism design problems. They have

showed that optimal noncollusive mechanisms can be made collusion-proof under a variety of circumstances. Dequiedt (2007) studied collusion in an auction with binary type spaces and discussed the conditions when collusion is efficient in this auction. Quesada (2005) studied collusion as an informed principal problem, and showed the optimal collusion-proof contract is asymmetric.

Most of the related studies have generally ignored competition in procurement. However in most procurements, multiple firms are involved and firms compete with each other and try to win the bid. So how will competition affect competitiveness? Rose-Ackerman (1996) concludes that increase in competitiveness will help to reduce corruption. However, Laffont and N´ Guessan (1999) used a three-tier agency model in which the supervisor only has partial information on the firms and the effect of the competition is characterized by the positive relationship between competitiveness and supervision effectiveness. They found competition could sometimes increase corruption. Celentani and Ganuza (2002) considered a procurement model with a competitive mechanism and studied the impact of competitiveness on equilibrium corruption. They also conclude that competition could increase corruption. Burguet and Che (Spring 2004) studied corruption in competitive procurement auction where there are two or more firms who are different in efficiency, and there is a corrupt supervisor who is willing to manipulate his evaluation of a firm's proposed contract in exchange for a bribe. They conclude that when supervisor has large manipulation power, in equilibrium the more efficient firm lose the bid with a positive probability.

Collusive bidding is another major problem in public procurement. Previous studies mainly used statistical method to detect firms' behavior that are inconsistent with behavior under competition. Porter and Zona (1993) and (1999) studied empirical data with suspicious bidders and detected anomalous bidding behaviors as evidence for bid rigging.

Kawai and Nakabayashi (2014) studied collusion among construction firms using a dataset covering construction projects procured by the Japanese government. They examined rebids and identified about 1,000 firms whose conduct is inconsistent with competitive behavior.

Detection of corruption is a very difficult task, mainly because the lack of data of direct evidence. Fazekas et al. (2013) suggested a indicator of grand corruption using information from administrative data. Yet the indicator is based on assumptions of corruption signals and lacks direct evidence.

Detection of financial fraud, such as transaction fraud, has implemented Machine Learning and Data Mining methods apart from statistical method. Among them, SVM (Support Vector Machine) is much used in fraud detection in credit card transactions. Hejazi and Singh (2012) compared the accuracy of different SVM kernel functions.

However, SVM method is rarely used in empirical studies in economics. Furthermore, there is little empirical study of fraud in federal procurement.

In this paper, I use game theory, machine learning and statistical methods to study fraud in public procurement. The paper is organized in six parts. First, I study a procurement game of firms and their strategies of fraud. Second, I introduce the method of SVM (Support Vector Machine). In the third section, I gives a description of the datasets I use. In the fourth section, I use the SVM method to study the data and get the detection result. Following the results of SVM, I use Logit Regression to study the data, and find out the relationship of fraud risk, competition, and performance monitor.

I find that theoretically if firms' types are close enough to each other, they tend to implement the same strategy regarding whether or not to engage in fraud.

Empirical results show that around 28% of contracts by DOD (Department of Defense) during

the fiscal year 2004 - 2010 are problematic compared to the 22% of a random sample of the full data and 18% of 2012 year DOT (Department of Transportation). This result is, to some extent, consistent with expectation, since military procurement is considered by many to be more problematic than most other public procuring fields.

Furthermore, my results also show that competition slightly increases fraud risk, but the effects are small. This result seems contradicting to intuition, but is pointed out by a few previous papers including Laffont and N´ Guessan (1999) and Celentani and Ganuza (2002).

However, the results also imply that under the conditions of full and open competition and performance monitoring, competition could reduce fraud risk. In addition, performance-based service contracts and flexible-price contracts have a higher risk of fraud.

2 The Game of Fraud in Procurement

2.1 The Setting of the Game

In this model the government is taken as exogenous and we only look at the N suppliers. Fraud means the company does not produce the quality-price level as promised or as what it should be under complete competition.

Define quality-price as $Q_i = \frac{q_i}{p_i}$, and the true quality-price the firm provided $\tilde{Q}_i = \frac{\tilde{q}_i}{p_i}$. When the contract-required or promised quality-price level Q_i is different from the true $\tilde{Q}_i$, we say it is fraud. Each firm has to announce its type θ_i which represent the firm's capacity or efficiency level of production. θ_i could be estimated based on a series of features of the company, such as the size and the profit, etc.

Assume in this game, firms can observe whether or not the competitors are engaged in fraud, but they do not report their competitors, either because they do not have the evidence or some other reasons.

Firm i's profit of producing true $\tilde{q}_i$ at price p_i is $\pi_i = p_i - c(\theta_i, \tilde{q}_i)$. For simplicity we assume in this game the contract price is given, and the firms compete in qualities. In this case, we can write profit as a function of quality-price level and type. $\pi(\theta_i, \tilde{Q}_i) = p_i - c(\theta_i, \tilde{q}_i)$. For the cases when firms compete in prices is similar, where Q_i is determined by p_i.

Suppose there is a cost $\tilde{c}$ if the firm decides to be engaged in fraud. This could be seen as bribery or cover-up cost for collusion, for instance, bid rigging. Assume the probability of being found out is ρ and the penalty is M.

Because the decision of strategies are based on the information of the types of all the firms, we assume that, for firm i under competition when there is no corruption, collusion,or any other fraud, the probability of wining the contract will be a function of i's own type and other firms' types: $\beta(\theta_i, \theta_{-i})$. When firm i and its competitors are engaged in fraud, the probability of wining the contract is another function of i's own type and other firms' types: $\alpha(\theta_i, \theta_{-i})$. The point is when firms are all engaged in collusion or corruption, the firm with higher ability has more power, thus it gets the contract with higher probability. When the game is symmetric, that is, $\theta_i = \theta_j$, $\forall i, j \leq N$, we have $\beta(\theta_i, \theta_{-i}) = \alpha(\theta_i, \theta_{-i}) = \frac{1}{N}$. Assume both probability functions are continuous and symmetric.

When firm i engage in fraud but its competitors do not engage in fraud, firm i will win the contract with probability 1, but will take on the risk of being revealed.

Firm i's strategy of whether or not to engage in fraud is represented by:

$$s_i = \begin{cases} 1, & \text{to engage in fraud} \\ 0, & \text{not to engage in fraud} \end{cases}$$

2.2 The Game with Two Firms

Now consider the two-player game, where two firms, $i = 1, 2$, competing to win the bid.

When firms take the strategies (s_1, s_2), the payments for the two firms are $R(s_1, s_2) = (R_1(s_1, s_2), R_2(s_1, s_2))$. Firm 1's payments $R_1(s_1, s_2)$ are as follows:

$$R_1(1,1) = \alpha_1(\theta_1,\theta_2)[\pi(\theta_1,\tilde{Q}_1) - \tilde{c}](1-\rho) - \rho M \tag{1}$$
$$R_1(1,0) = [\pi(\theta_1,\tilde{Q}_1) - \tilde{c}](1-\rho) - \rho M \tag{2}$$
$$R_1(0,0) = \beta_1(\theta_1,\theta_2)\pi(\theta_1,Q_1) \tag{3}$$
$$R_1(0,1) = 0 \tag{4}$$

$R_1(1,1$ is the payment for firm 1 when both firms engage in fraud. In this case the probability of firm i's winning the contract is $\alpha_1(\theta_1,\theta_2)$. It pays the fraud cost and also takes on the risk of being revealed.

$R_1(1,0)$ is the payment for firm 1 when firm 1 engages in fraud but firm 2 does not. In this case firm 1 wins the contract with probability 1, but will take on the risk of being revealed and fined.

$R_1(0,0)$ is the payment for firm 1 when neither of firms engages in fraud. In this case the probability of firm i's winning the contract is $\beta_1(\theta_1,\theta_2)$.

$R_1(0,1)$ is the payment for firm 1 when firm 2 engages in fraud but firm 1 does not. In this case firm 2 wins the contract with probability 1, and firm 1 gets zero.

Firm 2's payments are similar.

$$R_2(1,1) = \alpha_2(\theta_1,\theta_2)[\pi(\theta_2,\tilde{Q}_2) - \tilde{c}](1-\rho) - \rho M \tag{5}$$
$$R_2(1,0) = 0 \tag{6}$$
$$R_2(0,0) = \beta_2(\theta_1,\theta_2)\pi(\theta_2,Q_2) \tag{7}$$
$$R_2(0,1) = [\pi(\theta_2,\tilde{Q}_2) - \tilde{c}](1-\rho) - \rho M \tag{8}$$

We can write the game as:

(s_1,s_2)	$s_2=1$	$s_2=0$
$s_1=1$	$R(1,1)$	$R(1,0)$
$s_1=0$	$R(0,1)$	$R(0,0)$

Suppose the two firms' type are close enough to each other in the sense that for some ε, $\|\theta_1 - \theta_2\| < \varepsilon$. we can see there are two possible NE (Nash Equilibrium) in this game, (1, 1) and (0, 0).

First, if $\alpha_1(\theta_1,\theta_2)[\pi(\theta_1,\tilde{Q}_1) - \tilde{c}](1-\rho) - \rho M > 0$, because the two θ are very close, and functions are all continuous, then firm 2 must also have $\alpha_2(\theta_1,\theta_2)[\pi(\theta_2,\tilde{Q}_2) - \tilde{c}](1-\rho) - \rho M > 0$. Thus we can see at (1, 1) the two firms will not deviate, while (1, 0) and (0, 1) could not be NE. If $[\pi(\theta_1,\tilde{Q}_1) - \tilde{c}](1-\rho) - \rho M < \beta_1(\theta_1,\theta_2)\pi(\theta_1,Q_1$, then it will be the same for firm 2, and (0, 0) will also be a NE. If $[\pi(\theta_1,\tilde{Q}_1) - \tilde{c}](1-\rho) - \rho M > \beta_1(\theta_1,\theta_2)\pi(\theta_1,Q_1)$, then there is only one NE in this game.

Second, if $\alpha_1(\theta_1,\theta_2)[\pi(\theta_1,\tilde{Q}_1) - \tilde{c}](1-\rho) - \rho M > 0$, by similar arguments we can see there could be only one NE, (0, 0), in this game when $[\pi(\theta_i,\tilde{Q}_i) - \tilde{c}](1-\rho) - \rho M < \beta_i(\theta_1,\theta_2)\pi(\theta_i,Q_i)$.

If $[\pi(\theta_i,\tilde{Q}_i) - \tilde{c}](1-\rho) - \rho M > \beta_i(\theta_1,\theta_2)\pi(\theta_i,Q_i)$, then there is no pure-strategy NE in this game. Yet for mixed strategies we can also see that their mixed-strategy probabilities will be very close to each other, because mixed strategies are linear combinations of their pure strategies.

The arguments above could be easily extended to a N-player game. Given other firms' types and strategies we can compare any pair of firms in the same way as above.

From the arguments we have the proposition:

Proposition 1. When the firms' types are close enough to each other, they will tend to implement the same strategy.

In the long run for repeated games we can see that the firms with similar types will always display similar strategies. Hence they will display similar patterns. Because of this property, I use Support Vector Machine method to classify and detect fraud in the data.

In empirical studies we use the firms' known features to approximate its type. Assume the feature vector is δ_i, and the type is a function of feature vector $\theta_i = \varphi(\delta_i)$. Although we don't know what the type function φ is, we know that by continuity of the function, when feature vectors are close enough to each other, type θ_i are close to each other too. When the firms features are close enough to each other, they will also tend to implement the same strategy.

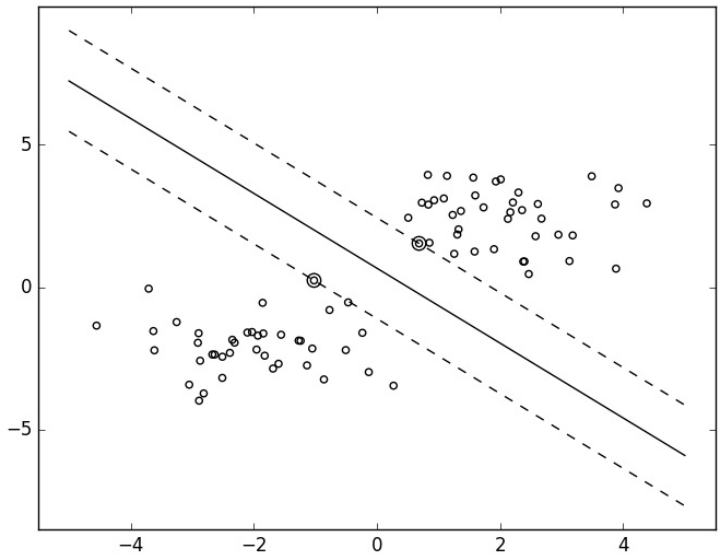

Figure 1: SVM Demonstration

3 Methodology

3.1 A Brief Introduction to Support Vector Machine

SVM (Support Vector Machine) has already been used in credit card fraud detection and online fraud transaction detection. The idea of this method is using program to investigate historical data of fraud and non-fraud and train a classifier, and then use the classifier to classify new data.

A linear SVM method with two classes of training data in a 2-dimensional Euclidean space is demonstrated by Figure 1.

For two classes of data points (x_i, y_i), where $x_i \in R^n$, $y_i \in \{-1, 1\}$. the objective of the method is to find the optimal boundary hyperplane between the two classes of data, which is the "most distant" from both groups.

The hyperplane is represented by $w \cdot x + b = 0$.

We need to make sure: $w \cdot x_i + b \geq 1$ when $y_i = 1$ and $w \cdot x_i + b \leq -1$ when $y_i = -1$.

Summarize these two conditions in one: $y(w \cdot x + b) \geq 1$

The margin of the hyperplane could be formulated as $\frac{2}{\|w\|^2}$, and while the problem is to maximize the margin, its dual problem is to minimize $\frac{\|w\|^2}{2}$

Because in most cases the two classes of data points are not separable, we need to allow for some space of mix. We add a parameter ξ to formulate the relaxed margin.

Thus the problem is:

$$\min_{w,b,\xi} \frac{\|w\|^2}{2} + C\sum_i \xi_i^k$$
$$\text{subject to } y(w \cdot x + b) \geq 1 - \xi_i^k.$$

Solve the minimization problem by First Order Condition using Lagrangian, we have the classifier function:

$$f(x) = sign(w \cdot x + b) = sign(\sum_i \lambda_i y_i x_i \cdot x + b) \tag{9}$$

For the non-linear cases where data points cannot be separated by linear hyperplane, we can use the kernel method. It maps the data points to a linear space, usually a higher-dimension space, by a function $\phi(x_i)$, and the problem can be formulated as:

$$\min_{w,b,\xi} \frac{\|w\|^2}{2} + C\sum_i \xi_i^k$$
$$\text{subject to } y(w \cdot \phi(x_i) + b) \geq 1 - \xi_i^k.$$

The classifier function becomes:

$$f(x) = sign(w \cdot \phi(x) + b) = sign(\sum_i \lambda_i y_i K(x, x_i) + b) \tag{10}$$

Where $K(x_i, x_j) = \phi(x_i)\phi(x_j)$ is called the kernel function. The problem is defined in this way so that we can use kernel functions directly in computer programming.

3.2 One-Class Support Vector Machine

Because of the lack of a controlled group of complete competitive data, I only use the problematic data to train a SVM classifier. In this case I implement the One-Class SVM method, which is often used to detect anomaly and to classify whether a new data point belong to a certain classification.

According to Schölkopf et al., the One-Class SVM problem can be formulated as:

$$\min_{w,b,\xi} \frac{\|w\|^2}{2} + \frac{1}{vN}\sum_i^N \xi_i - \tau$$
$$\text{subject to } w \cdot \phi(x_i) \geq \tau - \xi_i$$
$$\xi_i \geq 0$$

The classifier function in this case will be

$$f(x) = sign(w \cdot \phi(x) - \tau) = sign(\sum_i \lambda_i K(x, x_i) - \tau) \tag{11}$$

In the procurement game, if we can choose a kernel function such that ϕ resembles the type function of features $\varphi(\delta_i)$, then we can map the data points of features into the true type space, so that classifier could be more accurate.

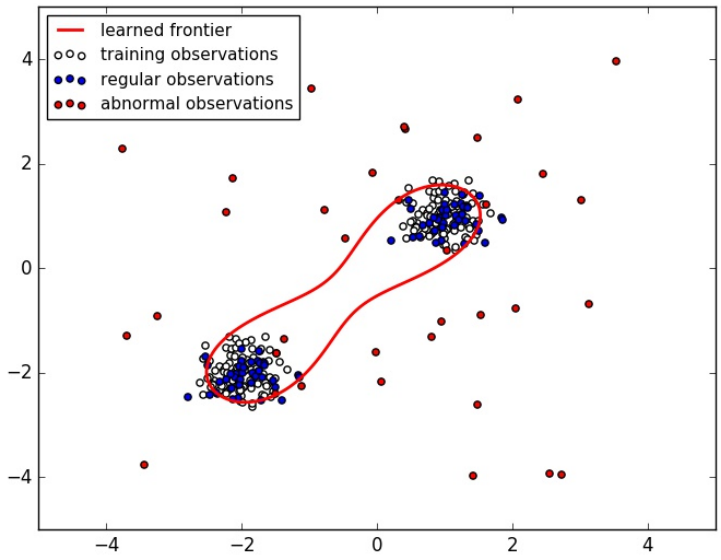

Figure 2: One-Class SVM with Non-linear Kernel

In my analysis, I used polynomial kernel function of degree 2:

$$K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j) = (\gamma(x_i \cdot x_j) + r)^2 \tag{12}$$

An example of One-Class SVM method with non-linear kernel function is demonstrated by figure 2.

The SVM program I use is the python Machine Learning package scikit-learn.

4 The Data

4.1 Federal Procurement Data System

The major data I use is the contract data from FPDS (Federal Procurement Data System) , downloaded from the data archives on usaspending.org.

The shortcomings of the data are: first, it does not have the other competitors' bids or proposals information, only the information about the winners of contracts. Second, the data is not cleaned. There are a lot of missing data and error inputs. GSA (General Services Administration) requires the Government agencies to collect and submit federal procurements data through FPDS. However, some of the Federal Government agencies did not report accurately, and some agents have problem understanding the FAR (Federal Acquisition Regulation) and how the system works.

The good thing about the data is that it has detailed information about all the federal government procurement contracts and the winning firms of contracts, such as number of employees, annual revenue of the firm, number of offers received for this contract etc. It also records the DUNS (Data Universal Numbering System) number of each contractor, so that we can track all the contracts won by a certain firm. Furthermore, the datasets are rich in size. The total data size of contracts from 2004 to 2013 is around 100 gigabytes. After screening, deleting errors, we still have considerably extensive datasets to work with.

Although the system also contains data from 2000 to 2003, it is considered to be low-quality, thus I only use data from 2004 and on. I mainly analyze contract data of DOD (Department of Defense) from 2004 to 2013, since DOD is the biggest spender among all the department. Military spending accounts for more than 50% of the budget. Moreover, Military spending is widely considered to be problematic.

I screened unreasonable inputs, dropped missing data and corrected some obvious errors.

In my study, I choose mostly variables that are numerical rather than categorical, for the sake of continuity.

4.2 Historical Exclusion and Misconduct Data

The additional information I use to construct datasets for training and testing are from SAM (System for Award Management) and FCMD (Federal Contractor Misconduct Database).

The System for Award Management provides a list of parties that are excluded from getting Federal procurement contracts because of misconduct, violation of some rules, or severe failure to fulfill responsibility. It records the types of exclusion by a Cause and Treatment code. I chose from the list the parties that are excluded because of fraud, corruption, violation of anti-trust laws. With their DUNS number, I collected their contract records in FPDS datasets of 2004-2009 and constructed a new dataset. This dataset contains the contracts won by these problematic firms before they were excluded by the government.

The test dataset I use is from an web database, FCMD (Federal Contractor Misconduct Database) of POGO (Project on Government Oversight). This database keeps the records of misconducts of top Fed contractors who are not excluded by the government, that is, they are not in the SAM exclusion list. I also chose the firms with misconducts in contract fraud, corruption, bid rigging. FCMD does not record their DUNS number, thus I searched their DUNS numbers online and make

up a list, and use the list to screen the FPDS datasets, and constructed a new dataset, similar to the SAM exclusion dataset.

The FCMD test dataset is more extensive, since these contractors are the top ones in Federal Procurements, and they are not excluded.

Both datasets are constructed from the FPDS datasets, so they have the same structure.

5 Data Analysis

5.1 Data Description

Description of the datasets for training, testing and detection are as follows:

Table 1: SAM Exclusion Data For Training

	contract value	# of offers	# of employees	revenue per employee
count	4116.000000	4116.000000	4116.000000	4.116000e+03
mean	164665.622796	3.088921	3291.324587	4.467484e+05
std	389778.268692	2.376344	1380.531794	2.335540e+06
min	-2429007.190000	0.000000	1.000000	3.750000e+02
25%	9413.750000	3.000000	3800.000000	4.473684e+05
50%	38538.000000	3.000000	3800.000000	4.473684e+05
75%	147271.000000	3.000000	3800.000000	4.473684e+05
max	7032650.000000	99.000000	32000.000000	1.500000e+08

Table 2: FCMD Data For Testing

	contract value	# of offers	# of employees	revenue per employee
count	5.069900e+05	506980.000000	506990.000000	5.069900e+05
mean	3.162380e+07	11.184595	19900.787986	2.790917e+06
std	3.547664e+09	83.705228	48792.932909	9.490613e+07
min	-1.920751e+08	0.000000	2.000000	5.933333e+00
25%	4.140000e+03	2.000000	2000.000000	2.186660e+05
50%	1.539400e+04	3.000000	2073.000000	3.376749e+05
75%	7.238689e+04	5.000000	12000.000000	2.086375e+06
max	2.430000e+12	999.000000	643000.000000	8.944444e+09

Table 3: 2004-2010 Department of Defense Data

	contract value	# of offers	# of employees	revenue per employee
count	3.381719e+06	3381719.000000	3.381719e+06	3.381719e+06
mean	1.025002e+08	4.693147	1.390037e+05	7.570933e+06
std	1.591615e+11	34.912529	1.436741e+07	6.798186e+09
min	-1.390192e+10	0.000000	1.000000e+00	5.709007e-05
25%	4.164960e+03	1.000000	2.200000e+01	1.179245e+05
50%	1.161500e+04	2.000000	1.590000e+02	2.272727e+05
75%	5.192500e+04	4.000000	3.800000e+03	4.500000e+05
max	2.923800e+14	999.000000	2.147484e+09	1.166667e+13

In the above tables, ”contract value” is the maximum value of three items in the dataset: ”dollars obligated”, ”base and all options value”, ”base and exercised options value”. The ”revenue per employee” is calculated by dividing ”annual revenue” by ”number of employees”.

I choose these variables because they are numerical and represent features of the firms and the contracts. Some other variables in the datasets I tried did not make a difference to the result, thus I abandoned them.

5.2 Testing and Detection

I use One-Class SVM with a polynomial kernel function to train the SAM Exclusion Data using python. The program developed a classifier function, and I use the classifier function to test the FCMD data and detect the 2004-2010 DOD (Department of Defense) data. In addition, I also test the classifier function on a random sample drawn from the contract datasets of the fiscal year 2004 to 2010, and the DOT (Department of Transportation) contract data of fiscal year 2012. The results are listed in the follow table:

Table 4: Testing and Detection of Four Datasets

	FCMD	Random Sample	2012 DOT	2004-2010 DOD
Count of -1	2.50928000e+05	7.65900000e+03	1.50810000e+04	2.43242700e+06
Count of 1	2.56052000e+05	2.19200000e+03	3.42600000e+03	9.49292000e+05
% of class 1	0.50505	0.222515	0.185119	0.28071

Here classification 1 represents the problematic contracts that have a higher fraud risk. Classification -1 represents the normal contracts. We can see the testing group, the FCMD data has approximately 50% of the contracts that are problematic. Compared to the results of the other two control groups, the random sample and the 2012 DOT, that both are approximately 20% problematic, the classifier does capture part of the problem we are dealing with.

Finally, the detection on the targeted dataset, the 2004-2010 year Department of Defense contract data, has a slightly higher ratio of problematic contracts, which is consistent with the expectation, because military procurement is less open and involved in more R&D products compared to other departments.

5.3 Competition, Performance Monitor, and Type of Contracts

With the classification result, I can now study the relationships of fraud risk and the independent variables, especially the relationship with competition level which is a much discussed top in theoretical studies, yet there is little evidence of empirical studies.

I replace the classification -1 with 0, and used Logit Regression of the classification result. The regression result is shown in the following table:

Table 5: Logit Regression Result

	coef	std err	z	P>|z|	[95.0% Conf. Int.]
contract value	3.081e-06	8.19e-09	376.216	0.000	3.06e-06 3.1e-06
extent competed	-0.1002	0.010	-10.509	0.000	-0.119 -0.082
# of offers	0.0009	0.000	7.455	0.000	0.001 0.001
# of employees	2.234e-07	6.43e-09	34.738	0.000	2.11e-07 2.36e-07
rev.per employee	3.008e-05	5.92e-08	508.116	0.000	3e-05 3.02e-05
performance based	0.2059	0.014	14.548	0.000	0.178 0.234
type of contract	0.1556	0.012	13.429	0.000	0.133 0.178
const	-13.6199	0.027	-512.856	0.000	-13.672 -13.568

Here I added three variables to the regression compared to the classification. In the data, the extent of competition is a categorical variable which has 8 categories. I transformed the extent of competition to a binary variable represented by the integers 0 and 1. Based on the degree of competition level, "Full and Open Competition", "Competed under SAP", "Full And Open Competition After Exclusion Of Sources" and "Competitive Delivery" are considered to be competitive, and are represented by 0. The others are all Non-Competitive procedures and are represented by 1.

The second variable is "performance based service contract" which is a binary variable, that is, it is either 0 or 1. If it is 1, then it is performance based service contract, then it will be evaluated and monitored by a set of rules made by the Federal Government including a work statement, measurable performance standards in terms of quality, timeliness, quantity, etc. and performance incentives where appropriate.

The third variable is "Type of Contracts" which means whether or not the contract is fixed-price. This variable includes multiple categories, yet only the "firm fixed-price" is represented by 0. The other categories are all flexible-price to some extent and are represented by 1.

In the above results we can see that competitiveness actually has a positive relationship with fraud risk, since when extent competed is 1, it is non-competitive. It implies the more competitive the procurement is, the more problematic the contract tends to be. The number of offers received also has a positive relationship with fraud risk. Some previous papers have already pointed out that competition might not reduce corruption, although this might contradict people's intuition. However, result shows that impact of the two variables are small.

The performance based service has a much bigger positive impact on fraud risk, which also contradicts the intuition. In this case, the interpretation could be that when there is performance based evaluation involved, there is larger space for rent seeking and bribery.

The positive coefficient of type of contract implies that flexible-price contracts are more problematic, which is consistent with expectation.

The table below demonstrates the results of logit regression on a subgroup of the performance based service and fully competitive contracts.

Table 6: Contracts of Full and Open Competition & Performanced Based

	coef	**std err**	**z**	**P>|z|**	**[95.0% Conf. Int.]**
contract value	3.96e-06	3.43e-08	115.460	0.000	3.89e-06 4.03e-06
# of offers	-0.0054	0.001	-4.309	0.000	-0.008 -0.003
# of employees	2.413e-07	2.33e-08	10.349	0.000	1.96e-07 2.87e-07
rev. per employee	2.806e-05	2.44e-07	114.877	0.000	2.76e-05 2.85e-05
const	-13.3583	0.107	-124.437	0.000	-13.569 -13.148

In this full and open competition and performance monitored subset, regression result shows that number of offers has a negative impact on fraud risk, that is, the more offers received, the lower the fraud risk is. It implies under proper competitive circumstances with performance monitor, competition could lower the fraud risk.

6 Remarks and Future Work

Fraud and corruption is very hard to detect based on the real-world data. In this paper I implemented Machine Learning method which is not often used in economics paper. Although the method cannot predict or detect fraud accurately, the result shows it does captures some anomaly pattern of the problematic contracts and gives a reasonable implication.

In the One-Class SVM training I used polynomial kernel function. However, other kernel functions might capture the features better that we can try in the future.

Apart from the dataset only containing the winners' information, one of the other shortcomings is the data is arranged by contracts, not firms. To get all the datasets re-arranged by the firms will require massive load of work, but could be the direction of the future work.

7 Conclusions

This paper studies the public procurement game, and developed a theory that firms of types that are close enough to each other will tend to implement the same strategy regarding whether or not to engage in fraud. Based on this proposition, I used One-Class Support Vector Machine method to train the historical data of firms which have engaged in fraud, and developed a classifier to make detection on DOD (Department of Defense) procurement data. The result shows that around 28% of contracts by DOD during the fiscal year 2004 - 2010 are problematic, that is, of a higher fraud risk, compared to the results 22% of a random sample and 18% of 2012 year DOT (Department of Transportation).

Further analysis of the classification result using Logit Regression shows that competition actually increases fraud risk, but the effects are small. However, under full and open competition and performance based monitoring, number of offers received has a negative relationship with fraud risk. In addition, the result also shows that performance based service contracts and flexible-price contracts are exposed to higher risk of fraud.

References

Emmanuelle Auriol. Corruption in procurement and public purchase. *International Journal of Industrial Organization*, 24:867C 885, Jan. 2006.

Edward C. Banfield. Corruption as a feature of governmental organization. *Journal of Law and Economics,*, 18(3):587–605, Dec. 1975.

Gary S. Becker and George J. Stigler. Law enforcement, malfeasance, and compensation of enforcers. *The Journal of Legal Studies*, 3(1):1–18, Jan. 1974.

Roberto Burguet and Yeon-Koo Che. Competitive procurement with corruption. *RAND Journal of Economics*, 35(1):50–68, Spring 2004.

Marco Celentani and Juan-José Ganuza. Corruption and competition in procurement. *European Economic Review*, 46:1273C1303, 2002.

Yeon-Koo Che and Jinwoo Kim. Robustly collusion-proof implementation. *Econometrica*, 74(4): 1063C1107, July 2006.

Yeon-Koo Che and Jinwoo Kim. Optimal collusion-proof auctions. *Journal of Economic Theory*, 144:565C603, 2009.

Vianney Dequiedt. Efficient collusion and optimal auctions,. *Journal of Economic Theory*, 136(1): 302–323, Sept. 2007.

Antoine Faure-Grimaud, Jean-Jacques Laffont, and David Martimort. Collusion, delegation and supervision with soft information. *Review of Economic Studies*, 70(2):253C279, April 2003.

Mihály Fazekas, István János Tóth, and Lawrence P King. Anatomy of grand corruption: A composite corruption risk index based on objective data. *Corruption Research Center Budapest Working Papers No. CRCB-WP/2013*, 2, 2013.

M Hejazi and YP Singh. Credit data fraud detection using kernel methods with support vector machine. *Journal of Advanced Computer Science and Technology Research*, 2(1):35–49, 2012.

Kei Kawai and Jun Nakabayashi. Detecting large-scale collusion in procurement auctions. *Available at SSRN 2467175*, 2014.

Fred Kofman and Jacques Lawarrée. Collusion in hierarchical agency. *Econometrica*, 61(3):629–656, May 1993.

Jean-Jacques Laffont and David Martimort. Collusion under asymmetric information. *Econometrica*, 65(4):875–911, Jul. 1997.

Jean-Jacques Laffont and David Martimort. Collusion and delegation. *The RAND Journal of Economics,*, 29(2):280–305, Summer 1998.

Jean-Jacques Laffont and David Martimort. Mechanism design with collusion and correlation. *Econometrica*, 68(2):309–342, Mar. 2000.

Jean-Jacques Laffont and Tchétché N´ Guessan. Competition and corruption in an agency relationship. *Journal of Development Economics*, 60:271–295, 1999.

Roger B. Myerson. Optimal auction design. *Mathematics of Operations Research*, 6(1):58–73, Feb. 1981.

OECD. *Bribery in Public Procurement:METHODS, ACTORS AND COUNTER-MEASURES*. OECD Publishing, 2007. ISBN 978-92-64-01394-0.

Trond E. Olsen and Gaute Torsvik. Collusion and renegotiation in hierarchies: A case of beneficial corruption. *International Economic Review*, 39(2):413–438, May 1998.

Robert Porter and John Zona. Detection of bid rigging in procurement auctions. *Journal of Political Economy*, 101(3):518–38, 1993.

Robert H Porter and J Douglas Zona. Ohio school milk markets: an analysis of bidding. *RAND Journal of Economics*, 30(2):263–288, 1999.

Lucía Quesada. Collusion as an informed principal problem. Oct. 2005. Unpublished Manuscript, University of Wisconsin.

Susan Rose-Ackerman. The economics of corruption. *Journal of Public Economics*, IV(1975): 187–203, 1975.

Susan Rose-Ackerman. Redesigning the state to fight corruption. (75), April 1996.

Bernhard Schölkopf, John C Platt, et al. Support vector method for novelty detection.

Andrei Shleifer and Robert W. Vishny. Corruption. *The Quarterly Journal of Economics*, 108(3): 599–617, Aug. 1993.

Jean Tirole. Hierarchies and bureaucracies: On the role of collusion in organizations. *Journal of Law, Economics, & Organization*, 2(2):181–214, Autumn 1986.