IDENTIFYING BID LEAKAGE IN PROCUREMENT AUCTIONS: MACHINE LEARNING APPROACH

DMITRY I. IVANOV AND ALEXANDER S. NESTEROV

ABSTRACT. We propose a novel machine-learning-based approach to detect bid leakage in first-price sealed-bid auctions. We extract and analyze the data on more than 1.4 million Russian procurement auctions between 2014 and 2018. As bid leakage in each particular auction is tacit, the direct classification is impossible. Instead, we reduce the problem of bid leakage detection to Positive-Unlabeled Classification. The key idea is to regard the losing participants as fair and the winners as possibly corrupted. This allows us to estimate the prior probability of bid leakage in the sample, as well as the posterior probability of bid leakage for each specific auction. We find that at least 16% of auctions are exposed to bid leakage. Bid leakage is more likely in auctions with a higher reserve price, lower number of bidders and lower price fall, and where the winning bid is received in the last hour before the deadline.

Keywords: corruption, bid leakage, procurement auctions, positive-unlabeled learning

JEL Classification: C38, C57, D82, H57

1. INTRODUCTION

In each country public procurement is an important and complex sector of the economy. In 2017 in Russia, the annual total volume of public procurement was 36.5 trillion rubles, which amounts to around a third of the annual GDP. A majority of contracts in Russian procurement are awarded through auctions, which in theory allocates the contract to the most efficient firm at the lowest possible price. In practice, however, certain tacit manipulations can corrupt the outcome both in terms of efficiency of allocation and the contract price.

In this paper we study “requests for quotations” – small and frequent online first price sealed-bid procurement auctions. These auctions can suffer from bid leakage – the corruption scheme where procurer illegally provides his favored participant with the information about the bids of the other participants. Our goal is to estimate how widespread bid leakage
is in general and to determine how likely it is that each particular auction has been affected by bid leakage.

We analyze the dataset containing more than 1.4 mln Russian requests for quotations. The dataset covers all the auctions that took place from January 2014 to March 2018 and is extracted from the online database.

![Figure 1. Example of typical request for quotations with leaked bids](ftp://ftp.zakupki.gov.ru/)

Notes: The auction lasts more than 7 days. The auction is suspicious for bid leakage as the winner bids near the deadline, after every other bid, and only slightly below the runner-up.

Our work is inspired by Andreyanov et al. (2016) who observed the patterns that are likely to reflect rational behavior of the favoured participant that received leaked bids (see Figure 1): these participants are

(1) bidding last,
(2) bidding close to the deadline, and
(3) winning by a small margin.

The intuition behind these three patterns in auctions with bid leakage is straightforward. First, the only way for the unfair participant to know every other bid and ensure her win is to bid the last, hence pattern (1). Similarly, she delays bidding as much as possible to lower the risks of not being the last, hence pattern (2). As she aims for the highest profit, she slightly undercuts the current best bid, hence pattern (3).

We use these three patterns to determine whether a particular auction has been corrupted by bid leakage. To do that we develop a two-stage identification strategy.

In the first stage we build a classifier that distinguishes the winners from the runner-ups by using features associated with patterns (1), (2) and (3). For a given auction winner, the higher is the predicted probability of winning, the more suspicious the auction is.

In a world without bid leakage, and assuming that these features are not related to the actual bid, such classifier would fail and if it does not then it has to be due to bid leakage. In practice however, the classifier might still be able to predict the winners well even in auctions without bid leakage, which leads to biased estimates. To correct our estimates we construct a synthetic placebo dataset of fair auctions and estimate the sign and the size of the bias.

![Figure 1. Example of typical request for quotations with leaked bids](ftp://ftp.zakupki.gov.ru/)
In the second stage we use the classifier’s predictions and performance to estimate the prior probability that a random auction in the dataset is corrupted, and the posterior probability that a specific auction is corrupted – conditional on the probability of winning that the classifier has assigned to its winner.

We estimate the prior probability of bid leakage as 16%. We also find that the bid leakage is more likely in auctions with a higher reserve price lower number of bidders and lower price fall, and where the winning bid is received in the last hour before the deadline.

The rest of the paper is organized as follows. In Section 2 we present the background on requests for quotations and the relevant literature and sketch our identification strategy. In Section 3 we describe the dataset. In Section 4 we present the two stages of our bid leakage estimation: the classifier and the estimates for prior and posterior probability of bid leakage. In Section 5 we present the results of estimation, provide few robustness checks and economic implications. Section 6 concludes.

2. Problem setup and identification strategy

2.1. Requests for quotations and background on bid leakage detection. In Russia requests for quotations are used for distributing small contracts such as roof repair for a factory or products delivery to a school kitchen. Before each auction starts, the procurer makes an announcement with the relevant information about the contract and the auction. The announcement includes reserve price – the maximal price the contract can be assigned for, the reserve price is bounded by 500000 rubles (approximately $8000). The auction lasts at least one week. During this period potential participants can submit their bids. Each participant can submit only one bid, the bids are sealed. After the auction ends, all bids are revealed and the smallest bid wins, the final price equals the winning bid (first-price auction).

Throughout the paper we only study successful bid leakage, that is when the honest winning bid has been leaked to and undercut by a favored bidder.

The literature on manipulations in auctions is prolific but mostly concerned with collusion schemes such as bid rigging Porter and Zona (1993) and bid rotation Aoyagi (2003); a rather recent review of the literature on collusion detection is available in Harrington (2005).

Crucial to our research question is the timing of bids, and the Russian procurement data is unique to contain this information. Previously, timing of bids has been studied in repeated Internet auctions such as eBay, where each bidder has a set of moments in time where he can submit a bid Song (2004). But, to the best of our knowledge, timing of bids has not been used before to detect corruption (except for Andreyanov et al. (2016) that we discuss below).
Other papers studying bid leakage or other forms of corruption using Russian data do so on a local scale of specific market or during specific period of time Yakovlev et al. (2016); Mironov and Zhuravskaya (2016); Balsevich and Podkolzina (2014). The empirical literature on auctions is rich (see, e.g., the seminal works Athey and Haile (2007) and Krasnokutskaya and Seim (2011)), yet there are only few studies that use supervised machine learning (classification) in auction corruption detection. Typically they utilize small datasets of few hundreds of labeled auctions Ferwerda et al. (2017); Huber et al. (2018).

2.2. Identification strategy and placebo. The only closely related papers to ours are Andreyanov et al. (2016) and its recent version Korovkin et al. (2018): we study the same object using the same data. However, our identification and estimation strategy is different in three crucial ways: we use weaker assumptions, more advanced methods and larger set of characteristics (i.e., features), which additionally enables us to determine posterior probability of bid leakage for each specific auction.

The identification in Andreyanov et al. (2016); Korovkin et al. (2018) relies on a crucial assumption that, in the auctions without bid leakage, bids and timing of the bids are independent. If independence holds, then the higher likelihood of the last bids to win compared to earlier bids is attributed exclusively to bid leakage.

The independence assumption might fail for a number of reasons. The longer time it takes a risk-averse bidder to study the case, the lower will his bid be. For example, in our data we observe that 1,2% of participants behave like “snipers”: they bid during the first day of the auction and bid slightly below the reserve price (up to 5%).

Another reason comes from the honest bidders’ attempts to resist bid leakage. On the one hand, the later the bid is submitted, the lower is the chance that it is going to be leaked and undercut by a corrupted bidder. On the other hand, submitting closer to the deadline requires attention and possibly costly logistics. As a result, a bidder with a higher valuation (i.e., lower execution costs) submits a lower bid and, simultaneously, is ready to delay the submission more relative to a bidders with a lower valuation. Because bidding and timing are confounded through valuation, the independence assumption does not hold. A later bid has a higher chance of winning not only due to presence of bid leakage, but due to mere expectation of bid leakage. We present this argument formally in the Subsection 2.3.

We do not rely on independence assumption. Instead, we correct our estimates using a synthetic placebo dataset of fair auctions. We remove the first-ranked bidders (the true winners) from all the auctions and recalculate the features accordingly. This way, we obtain a new dataset where the second places are treated as the winners and the third places – as the runner-ups (Figure 2). We estimate the bias in these synthetic auctions and assume

\[ \text{Anecdotal evidence suggests that some bidders did not rely on post or courier services and delivered their bids personally to make sure to submit just before the deadline.} \]
that this bias is equal to the bias in fair auctions in the original dataset. We verify this assumption and compare it to independence in Section 4.3.

| Announcement | Deadline |
|--------------|----------|
| Reserve Price | Bid 1 | Bid 2 | Computed Bid |
| 150000 R     | 120000 R | 145000 R | 119500 R |

**Figure 2.** Placebo auction example, transformed from the auction at Figure 1

Notes: The placebo auction is generated by removing the winning bid from the auction. In this hypothetical auction, bid 1 belongs to the winner, bid 2 – to the runner-up. The auction is no longer suspicious for bid leakage.

Second, in contrast to reduced-form and structural statistical models we use machine learning techniques. This allows us to consider all the evidence on bid leakage patterns at once and without imposing restrictions on their interconnection. When we only include subset of the relevant features as it is done in [Korovkin et al. (2018)](#), we get significantly less accurate predictions.

Finally, we descend from the population level and develop a method able to assign the posterior probability of the bid leakage presence to every auction in the dataset. Our method provides more precise and specific estimates of bid leakage and can be used for automatic ex-post bid leakage detection, which can be useful for regulation and auditing authorities.

2.3. **Game-theoretic model of bid leakage.** In this Section we formalize the intuition behind our identification strategy using a simple game-theoretic model.

First consider the world without bid leakage. An auctioneer is selling a procurement contract with reserve price normalized to 1, and the lowest possible cost of executing the contract is normalized to zero. Each bidder is risk-neutral and is drawing his execution costs \( e \) from a uniform distribution on \([0, 1]\), or, equivalently, each bidder has an iid valuation \( v = 1 - e \) drawn from a uniform distribution on \([0, 1]\); the cumulative density function is \( F(v) = v \).

Let the expected number of bidders participating in the auction be \( n \). For each bidder \( i \) with a valuation \( v \) his equilibrium bid is the expected bid of his runner-up conditional on \( i \) being the winner,

\[
b^*(v) = 1 - \frac{\int_0^v x dF^{n-1}(x)}{\int_0^v dF^{n-1}(x)} = 1 - \frac{n - 1}{n}.
\]

\( ^3 \)Since bidders are risk-neutral and valuations are i.i.d. the uncertainty regarding the exact number of bidders does not play a role as shown by [Matthews (1987)](#) [McAfee and McMillan (1987)](#).
Since the bid is monotonic in the valuation, in order to win the bidder needs to have the highest valuation. The probability of winning is thus $F_{n-1} = v_n^{-1}$, and the expected profit is

$$E\pi(v, b^*(v)) = \frac{v_n}{n}.$$  

Now we add the time dimension to the problem. Each bidder chooses the submission time $t \in [0, 1]$. We assume that delaying submission is costly, submitting at time $t$ costs the bidder $c(t)$, where $c$ is increasing and convex, $c' > 0$, $c'' > 0$, and extremely high close to deadline, $c(1) = \infty$. These costs represent the stress and attention costs of not missing the deadline and also the costs of more precise bid delivery. In the world without bid leakage the timing of the bid is irrelevant for winning and each bidder submits at time $t = 0$.

Now let bid leakage be possible. We assume that each bidder has the same prior belief regarding the possibility of bid leakage. Conditional on that the auction is corrupted, the probability that a specific bid is leaked and undercut decreases in time of submission: if you submit later, then the chances of leakage of your bid are lower. We assume that for each bidder the perceived probability that his bid submitted at time $t$ is leaked and undercut is exogenously given by some function $\beta(t)$ decreasing in time, $\beta'(t) < 0$, down to zero at the time of deadline $t = 1$, $\beta(1) = 0$.

Thus the expected profit of the bidder with valuation $v$ and bid $b^*(v)$ is as follows:

$$E\pi(v, b^*(v), t) = \frac{v_n}{n}(1 - \beta(t)) - c(t).$$

The optimal submission time $t^*(v)$ is given by the first order condition:

$$c'(t^*(v)) = -\frac{v_n}{n}\beta'(t^*(v)).$$

Observe that both $b^*(v)$ is decreasing and $t^*(v)$ is increasing in valuation $v$. Thus the optimal bid and the submission time of the bid are confounded by the valuation. The higher the valuation is, the more is the bidder ready to pay to get the contract: both in terms of submitting a lower bid and in terms of costly delay of the submission.

Observe also that the correlation between the timing and the bid holds for each valuation, and thus will be true not only for the winners but also for the runner-ups. We uncover this correlation for runner-ups using a placebo dataset in Section 4.2 and use it to correct our biased estimates for the winners.

2.4. Positive-Unlabeled Classification. Our bid leakage estimation strategy is based on Positive-Unlabeled (PU) Classification. Generally, PU Classification is applied instead of Supervised Classification in the cases when the training data set is not fully labeled. Specifically, only a subsample of the Positive data needs to be labeled as such, while the remaining part of the Positive data and all the Negative data are mixed in the Unlabeled sample. This
Table 1. Data set characteristics

| Characteristics         | Mean  | Median | Std. Dev. |
|-------------------------|-------|--------|-----------|
| Number of participants  | 3     | 2      | 1.9       |
| Reserve price, rubles   | 182000| 134000 | 150000    |
| Winner’s bid, rubles    | 142000| 97000  | 128000    |
| Runner-up’s bid, rubles | 154000| 108000 | 134000    |
| Price fall              | 0.23  | 0.18   | 0.21      |
| Time from bid to deadline, hours | 40   | 20     | 54        |
| Time from winner’s bid to deadline, hours | 39  | 19     | 54        |
| Time from runner-up’s bid to deadline, hours | 39  | 20     | 53        |
| Duration, hours         | 195   | 169    | 71        |

setting may be applied to our case if we regard runner-ups as fair (Positive) and the winners as possibly corrupted (Unlabeled).

Numerous methods are proposed to solve PU Classification [Elkan and Noto (2008); Kiryo et al. (2017); Ivanov (2019)]. They are applied to various real-world problems, which include detection of fake texts [Ren et al. (2014)], time-series classification [Nguyen et al. (2011)], bioinformatics [Yang et al. (2012)], etc.

3. Auction Data

We extracted data\(^4\) on 1444718 requests for quotations that took place between January 2014 and March 2018. The data was preprocessed in the following way:

- The auctions with missing data or with obvious coding mistakes were dropped. These obvious mistakes include: reserve price being negative or higher than the upper bound 500000 rubles; the starting date being after the ending date; the starting date, the ending date or the bidding date being in the future; the bid being negative or higher than the reserve price. 86% of the initial size is left.
- Our identification methods cannot be applied to auctions with 1 participant, so these auctions are dropped. 44% of initial size is left.

The data on 636866 auctions remains after the preprocessing. The main characteristics of this data set are shown in Table 1.

Notes: The data set excludes auctions with 1 participant and auctions with missing data. We define Price fall as $\frac{r - b_1}{r}$, where $r$ is reserve price, $b_1$ is winner’s bid.

\(^4\)The procurement auctions’ data are stored at ftp://zakupki.gov.ru
Here we describe our two-stage bid leakage estimation strategy. Mainly, we reduce the problem to Positive-Unlabeled Classification by considering the runner-ups as fair (Positive) participants and the winners as a mixture of fair (Positive) and corrupted (Negative) participants.

We follow the state of the art DEDPUL procedure proposed in Ivanov (2019) for general purposes. At the first stage of the procedure a supervised binary classifier is trained to distinguish the winners from the runner-ups. Using cross-validation technique, predictions of this classifier are obtained for all the winners in the data set. At the second stage these predictions are transformed into bid leakage probabilities by estimating the ratio between the densities of the predictions for the runner-ups and for the winners.

There is a crucial distinction of our strategy compared to the original DEDPUL procedure. The original procedure would assume that bid leakage is the only reason why the winners and the runner-ups differ for the classifier. However, as we have already discussed in Sections 2.2 and 2.3, this might not be the case, and the difference may exist even in the fair auctions. To account for this we introduce into the analysis the synthetic placebo auctions defined in Section 2.2, which are assumed to be fair, and modify the procedure correspondingly. These modifications are discussed in details in Section 4.2.

### 4.1. First Stage: Winner vs Runner-up Classifier.

In the first stage we train the classifier to distinguish the bids of the winners from the bids of the runner-ups.

The features are presented in Table 2. These features are specifically designed to reflect possible bid leakage patterns, while uncovering only little information about fair auctions. Specifically, the features `bid last?` and `bid timing` reflect intention of a corrupted participant to gather information about all the other bids. Small values of `relative bid` reflect undercutting. The feature `met before?` reflects the possibility of repeated procurer-participant cooperation. Small values of `relative bid timing` might reflect fairness of participant, as bids are unlikely to be leaked instantly.

Notes: `relative bid` is truncated at 0.1: values bigger than 0.1 are set to this threshold. Likewise, `bid timing` and `relative bid timing` are truncated at 1440 minutes (1 day). Auctions with 1 participant are excluded from the analysis.

Observe that the information on whether some two participants are from the same auction is lost on purpose. The classifier does not choose the winner between the two participants in each auction. Instead, it determines the chances that each set of features in the data set belongs to a winner as opposed to a runner-up.
Table 2. Features description

| Name           | Type     | Range     | Description                                                                 |
|----------------|----------|-----------|-----------------------------------------------------------------------------|
| bid last?      | Binary   | {0, 1}    | Did participant bid after other participants?                              |
| met before?    | Binary   | {0, 1}    | Was participant in auction with this procurer before?                       |
| bid timing     | Continuous | [0, 1440] | Minutes from the moment bid is made to deadline                             |
| relative bid   | Continuous | [0, 0.1]  | Difference with bid of succeeding place, normalized by reserve price         |
| relative bid timing | Continuous | [0, 1440] | Difference in minutes with bid timing of previous minimal bid               |
| number of participants | Integer   | [2, 86]   | Number of participants in auction                                           |

As a classifier we use gradient boosting of decision trees – namely, xgboost[7]. We train an ensemble of 60 trees on the features described above with depth of each tree limited to 5 levels. With this classifier we obtain predicted probability of winning for each winner in the data set using cross-validation. At the second stage we establish the connection between these predictions and the probability of bid leakage.

4.2. Second Stage: Transforming Classifier’s Predictions into Bid Leakage Probability. We show how to use the discussed classifier to estimate both the prior and the posterior probabilities of bid leakage, neither of which are assumed to be known in advance for any of the auctions. First we introduce notations and formally define the problem.

At the first stage the classifier estimates the probability that the participant wins based on corresponding vector of features $x$. Denote this probability of winning as $y(x)$. Denote distributions of $y(x)$ for winners and runner-ups as $f_w(y)$ and $f_{\pi}(y)$ respectively.

As was previously discussed, we consider the runner-ups to be fair participants (as we aim to detect only successful bid leakage), while the winners may contain both fair and corrupted participants. Moreover, the distributions of $y(x)$ for the winners and the runner-ups of the fair auctions might also differ. This may generally be expressed in the following mixture model:

\begin{equation}
\label{eq:1}
f_{\pi}(y) = f_{corr_2}(y)
\end{equation}

\begin{equation}
\label{eq:2}
f_w(y) = \alpha f_{corr}(y) + (1 - \alpha)(f_{corr_2}(y) + \Delta_{12})
\end{equation}

[7]https://github.com/dmlc/xgboost
(3) \[ \Delta_{12} = f_{corr_1}(y) - f_{corr_2}(y) \]

where \( \alpha \) denotes the prior probability of bid leakage; \( f_{corr}(y) \), \( f_{corr_1}(y) \), and \( f_{corr_2}(y) \) denote the distributions of \( y(x) \) for the corrupted winners, the fair winners, and the fair runner-ups respectively; \( \Delta_{12} \) denotes the difference between the distributions of the winners and the runner-ups in fair auctions. Introduction of \( \Delta_{12} \) into (2) is exactly what distinguishes our case from the standard PU Classification problem setup. However, we will address the issue of estimating \( \Delta_{12} \) later. For now, consider it exogenous.

Our goal is to estimate the prior probability that a random winner is corrupted \( \alpha \) and the posterior probability that a specific winner is corrupted \( f(corr \mid y) \). The latter may be expressed using the Bayes rule:

(4) \[ f(corr \mid y) = \frac{\alpha f_{corr}(y)}{f_w(y)} = 1 - \frac{(1 - \alpha)(f_{\pi}(y) + \Delta_{12})}{f_w(y)} \]

Following DEDPUL (Ivanov, 2019), the densities \( f_w(y) \) and \( f_{\pi}(y) \) may be estimated by applying Kernel Density Estimation to the classifier’s predictions for the winners and the runner-ups respectively. Then, both priors \( \alpha \) and posteriors \( f(corr \mid y) \) may simultaneously be estimated by applying Expectation-Maximization algorithm to (4), thus reaching our goal.

Now we address the issue of \( \Delta_{12} \) estimation. The key step is to construct the synthetic data set of implicitly fair placebo auctions. As was previously discussed, placebo auctions are generated from the real auctions by dropping the winners and keeping the other participants which we know to be fair. In each hypothetical auction, the second-ranked bidder is assumed to be the winner, and the third-ranked bidder is assumed to be the runner-up.

By applying the classifier that is trained on the real auctions to the placebo data set, we may obtain its predictions to later estimate the densities \( f_{wp}(x) \) and \( f_{\pi_p}(x) \) for the winners and the runner-ups of placebo auctions respectively, where:

(5) \[ \Delta_{23} = f_{wp}(y) - f_{\pi_p}(y) \]

Thus, we may estimate \( \Delta_{23} \) by using placebo auctions. This becomes crucial as we make a major assumption regarding equality of \( \Delta_{12} \) and \( \Delta_{23} \):

**PARITY**: \( \Delta_{12} = \Delta_{23} \). The difference between the winners and the runner-ups in the real fair auctions is equal to this difference in the placebo auctions.
Using the parity assumption we may estimate $\Delta_{12}$ and thus the priors $\alpha$ and the posteriors $f(corr \mid y)$ of bid leakage. In the next subsection we verify applicability of the parity assumption.

4.3. Verifying parity and independence assumptions. First observe that independence implies parity. Namely, if timing is independent from bidding, then both differences $\Delta_{12}$ and $\Delta_{23}$ (when estimated using only time-related features) are equal to zero $\Delta_{12} = 0 = \Delta_{23}$ implying parity.

To test the independence assumption exclude all winning last bids and then test if being last predicts a lower bid. They find the opposite effect (in this subsample earlier bids are lower) of a much lower size. We replicate this observation with our dataset.

However, if we exclude all winning bids (and not only those bids that were submitted last as in Korovkin et al. (2018)), we find that later bids are more likely to be smaller. That is, in the imaginary auction with only 2nd and 3rd lowest bids, the lower 2nd bid is on average submitted later. This is also true when we exclude winners and 2nd lowest bids, and compare 3rd and 4th latest bids, the lower 3rd bid is on average submitted later.

Both independence and parity assumptions may be tested by training the classifier on the real auctions and applying it to the placebo auctions. In the case if independence holds, we expect the classifier’s performance on the placebo data set to be on a level of fair coin (0.5 accuracy and ROC-AUC), which is clearly not the case (Column 2 in Table 3). In the case if parity holds, we expect the classifier to show the same performance on the placebo data sets regardless of how many participants are dropped. Observe that this is the case for two placebo data sets: when only the winners are dropped (Column 2 in Table 3) and when both winners and runner-ups are dropped (Column 3 in Table 3). At the same time, classifier’s performance on the real data set (Column 1 in Table 3) is considerably higher than on the placebo data sets, which is expected in the presence of bid leakage.

Thus the evidence suggests that the parity assumption holds, meaning that our bid leakage estimation strategy is applicable, while the stronger assumption of independence does not. Both results hold when the bid-related features are excluded.

5. Empirical results

The overall prior probability of bid leakage in our sample is estimated as 16%, which is slightly higher than 10-11% found previously in Korovkin et al. (2018). One of the reasons for this discrepancy is that our sample also includes auctions with only two participants, and bid leakage is significantly more likely there. The other reasons are due differences in identification and estimation.
Table 3. Measures of classifier’s performance on different data sets: on real auctions (first column), on placebo auctions with dropped winners (second column), on placebo auctions with dropped winners and runner-ups (third column)

|                  | $\Delta_{12}; f_{corr}$ | $\Delta_{23}$ | $\Delta_{34}$ |
|------------------|--------------------------|---------------|---------------|
| Accuracy         | 0.5819 +/- 0.0005         | 0.5375 +/- 0.0008 | 0.5387 +/- 0.0012 |
| ROC-AUC          | 0.6245 +/- 0.0004         | 0.5581 +/- 0.0008 | 0.5534 +/- 0.0011 |

Notes: mean and standard deviation statistics of scores are calculated on outputs of 3-fold cross-validation, repeated several times. Since classifier’s purpose is to learn the patterns that reflect bid leakage, its suitability for our estimation strategy should be evaluated with the difference of the scores on the real (first column) and the placebo (second and third columns) data sets, rather than solely the score on the real data set.

The probability $\alpha$ for specific subsamples provides insights into the mechanisms behind bid leakage. We present these results in Fig.3.

Figure 3. Bid leakage probability aggregated by auction characteristics

The top left diagram in Fig.3 demonstrates that as the reserve price increases, the bid leakage is more likely to be observed. This is very natural and might have two explanations. First, a higher reserve price for a contract gives higher incentives to organize risky corruption schemes. Second, the better organized corruption schemes allow setting a higher reserve price in order to maximize the surplus.

The top center diagram in Fig.3 demonstrates that the lower number of bidders on average corresponds to higher probability of bid leakage. The number of participants in an auction is endogenous: the entry is always costly and the expected benefit depends on the reserve price, description of the contract and other details posted in the announcement. If some of these details (such as required certification) deter entry or signal the experienced participants that
bid leakage is likely, which results in fewer bidders entering the auction. If there is no such deterrence, then the auction is desirable and entry is high. And a desirable auction is more likely to be the one where bid leakage occurs.

As the number of bidders grows above four, the relation to bid leakage becomes ambiguous. On the one hand, higher participation might mean that the auction was more competitive, which corresponds to low $\alpha$. On the other hand, as demonstrated earlier bid leakage is correlated with reserve price, and an overly high reserve price might attract more bidders.

The top right diagram in Fig.3 demonstrates that there is no particular pattern in seasonality of bid leakage. The only exception is December, where the procurement agencies might be distributing their accumulated reserves.

The bottom left diagram in Fig.3 shows a stark relation of bid leakage and the difference between the final price and the reserve price. When surplus is high, price fall can be a measure of bidders’ competition. In general, the larger is the price fall, the lower is the probability of bid leakage. However, if the price fall is below 5% then the predicted probability of bid leakage increases with the price fall. This can be due to that the reserve price has been already set very close to bidders’ valuations, providing little opportunity for bid leakage.

The bottom center diagram in Fig.3 demonstrates that commission size has very low influence on the probability of bid leakage. There is however a marginally significant difference between commissions with 3 members and commission with 7 members, which naturally suggests that larger commissions correspond to lower bid leakage.

The bottom right diagram in Fig.3 shows that the (estimated) favored participants submit their bids at constant rates throughout the last 24 hours of the auction, except for the last hour, when the submission rate is 2-3 times higher.

Finally, Fig.4 presents the cross-regional average probabilities of bid leakage. Perhaps surprisingly, the variance between different regions is rather large, up to 90%. However, these results are consistent with other measures of corruption available for Russian regions, such as electoral fraud (see, e.g., Mebane and Kalinin (2009), Bader and van Ham (2015)).

6. Conclusions

We study first-price sealed-bid auctions and identify auctions corrupted with bid leakage. The first stage of our strategy is to build a classifier that very well distinguishes the winners from the runner-ups in the corrupted auctions but not as well in the fair auctions. In the second stage we process the classifier’s predicted probabilities of winning for the winner and the runner-up of each auction into the probability that the auction is corrupted. We apply our estimation strategy to the Russian procurement data between January 2014 and March 2018 containing 636866 auctions.
We estimate the share of the corrupted auctions in the dataset as 16%. We believe that this estimate is conservative due to the assumptions we make. First, we are only concerned with effective bid leakage, that is if the bids are leaked to the favored participant, she inevitably wins. Consequently, when the classifier selects a runner-up, we assume it to be a mistake. Second, we assume that the classifier selects all the winners of the corrupted auctions. Consequently, when the classifier doesn’t select the winner of an auction, we assume this auction to be fair.

Several problems remain open. First, our current strategy can only be applied to the auctions with 2 and more participants. Second, and arguably the most important problem is that the validation of our strategy in its classical way of comparison with labeled data is not available. Yet the confidence in our strategy is reinforced by the economic interpretability of the results: the bid leakage is more likely in auctions with a higher reserve price, lower number of bidders and lower price fall, and where the winning bid is received in the last hour before the deadline.

REFERENCES

Pasha Andreyanov, Alec Davidson, and Vasily Korovkin. 2016. *Corruption vs Collusion: Evidence from Russian Procurement Auctions*. Technical Report. mimeo: UCLA.
Masaki Aoyagi. 2003. Bid rotation and collusion in repeated auctions. *Journal of economic Theory* 112, 1 (2003), 79–105.

Susan Athey and Philip A Haile. 2007. Nonparametric approaches to auctions. *Handbook of econometrics* 6 (2007), 3847–3965.

Max Bader and Carolien van Ham. 2015. What explains regional variation in election fraud? Evidence from Russia: a research note. *Post-Soviet Affairs* 31, 6 (2015), 514–528.

Anna Balsevich and Elena Podkolzina. 2014. Indicators of Corruption in Public Procurement: The Example of Russian Regions. (2014).

Charles Elkan and Keith Noto. 2008. Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 213–220.

Joras Ferwerda, Ioana Deleanu, and Brigitte Unger. 2017. Corruption in public procurement: finding the right indicators. *European Journal on Criminal Policy and Research* 23, 2 (2017), 245–267.

Joseph E Harrington. 2005. *Detecting cartels*. Technical Report. Working Papers, The Johns Hopkins University, Department of Economics.

Martin Huber, David Imhof, et al. 2018. *Machine Learning with Screens for Detecting Bid-Rigging Cartels*. Technical Report. Faculty of Economics and Social Sciences, University of Freiburg/Fribourg . . .

Dmitry Ivanov. 2019. DEDPUL: Method for Mixture Proportion Estimation and Positive-Unlabeled Classification based on Density Estimation. *arXiv preprint arXiv:1902.06965* (2019).

Ryuichi Kiryo, Gang Niu, Marthinus C du Plessis, and Masashi Sugiyama. 2017. Positive-unlabeled learning with non-negative risk estimator. In *Advances in Neural Information Processing Systems*. 1675–1685.

Vasily Korovkin, Pasha Andreyanov, and Alec Davidson. 2018. Detecting Auctioneer Corruption: Evidence from Russian Procurement Auctions. (2018).

Elena Krasnokutskaya and Katja Seim. 2011. Bid preference programs and participation in highway procurement auctions. *American Economic Review* 101, 6 (2011), 2653–86.

Steven Matthews. 1987. Comparing auctions for risk averse buyers: A buyer’s point of view. *Econometrica: Journal of the Econometric Society* (1987), 633–646.

R Preston McAfee and John McMillan. 1987. Auctions with a stochastic number of bidders. *Journal of economic theory* 43, 1 (1987), 1–19.

Walter R Mebane and Kirill Kalinin. 2009. Comparative election fraud detection. (2009).

Maxim Mironov and Ekaterina Zhuravskaya. 2016. Corruption in procurement and the political cycle in tunneling: Evidence from financial transactions data. *American Economic Journal: Economic Policy* 8, 2 (2016), 287–321.
Minh Nhut Nguyen, Xiaoli-Li Li, and See-Kiong Ng. 2011. Positive unlabeled leaning for time series classification. In IJCAI, Vol. 11. 1421–1426.

Robert H Porter and J Douglas Zona. 1993. Detection of bid rigging in procurement auctions. Journal of political economy 101, 3 (1993), 518–538.

Yafeng Ren, Donghong Ji, and Hongbin Zhang. 2014. Positive Unlabeled Learning for Deceptive Reviews Detection. In EMNLP. 488–498.

Unjy Song. 2004. Nonparametric estimation of an eBay auction model with an unknown number of bidders. (2004).

Andrei Yakovlev, Oleg Vyglovsky, Olga Demidova, Alexander Bashlyk, et al. 2016. Incentives for repeated contracts in public sector: empirical study of gasoline procurement in Russia. International Journal of Procurement Management 9, 3 (2016), 272–289.

Peng Yang, Xiao-Li Li, Jian-Ping Mei, Chee-Keong Kwoh, and See-Kiong Ng. 2012. Positive-unlabeled learning for disease gene identification. Bioinformatics 28, 20 (2012), 2640–2647.