Intl. Trans. in Op. Res. xx (2021) 1–25

Explainable anomaly detection for procurement fraud identification - lessons from practical deployments

Adam Westerski[a,*], Rajaraman Kanagasabai[a], Eran Shaham[a], Amudha Narayanan[a], Jiayu Wong[b], Manjeet Singh[b]

[a]*Institute for Infocomm Research, A*STAR, Singapore*
[b]*A*STAR Procurement Office, A*STAR, Singapore*
E-mail: adam-westerski@i2r.a-star.edu.sg [Westerski]; kanagasa@i2r.a-star.edu.sg [Kanagasabai]; eran-shaham@i2r.a-star.edu.sg [Shaham]; naraa@i2r.a-star.edu.sg [Narayanan]; wong_jiayu@hq.a-star.edu.sg [Wong]; manjeet_Singh@artc.a-star.edu.sg [Singh]

Received DD MMMM YYYY; received in revised form DD MMMM YYYY; accepted DD MMMM YYYY

Abstract

This article reports on results of our work to construct a system for detection of fraudulent behaviour in procurement transactions. To solve the problem, we model different types of fraud via separate statistical indicators. We propose a formalised framework to describe the severity of fraud in a unified way regardless of underlying fraud mechanics. Subsequently, we leverage this concept to build indicator ensembles which collect evidence from multiple indicators and deliver an interpretable per transaction score to the procurement audit officer. As a case study, we overview 48 of such fraud indicators constructed for our client and describe two examples in detail showing how our formal definitions can be transformed into a practical implementation. The presented results include experiments with all indicators on data covering four years of procurement activity with approximately 216,000 transactions coming from a large government organisation in Singapore. The final evaluation of our system shows 67.1% precision in detecting suspicious transactions. The article describes how outcome of our work helped to effectively cope with problem of anomaly detection explainability and the lessons learned from integrating this solution to operational practices of a procurement department.

Keywords: Procurement; Data Analytics; Unsupervised; Fraud Detection; Anomaly Detection; Explainable AI

1. Introduction

For many organisations procurement is an important element of managing their operations. The value of procured goods can be a significant portion of spending, up to 60% revenue depending on the industry (Bartolini, 2011). Therefore, companies increasingly invest in optimising procurement operations by

*Author to whom all correspondence should be addressed (e-mail: adam-westerski@i2r.a-star.edu.sg).

reorganising their practices to root out inefficiencies and problems (Grand View Research, 2019). Typically, the savings are made through consolidation of purchases, closer collaboration with vendors and promoting competition during contract bidding (Umbenhauer and Gregson, 2016). However, beyond such measures, lack of adequate procurement management can cause organisation loss due to misdetection of more intensional harmful practices, ie. procurement fraud (Supply Management. Chartered Institute of Procurement & Supply (CIPS), 2014; PwC, 2014). Such type of fraud can occur due to: external vendors exploiting gaps in organisation procurement procedures; employees acting as purchase requesters taking advantage of internal purchase procedures for their own benefit; or potentially collusion between those two parties: vendors and requesters working together to capitalise at the expense of the organisation they exploit. While such instances of fraud are difficult to detect, they can be very harmful in terms of monetary loss as well as severe damage to organisation reputation.

In academic literature, there have been multiple attempts to solve problems of procurement management, including procurement fraud (Ruping et al., 2008; Calafato et al., 2014; Jans et al., 2010). Many of such works have been based on extensive theoretical frameworks that assume existence of data with certain features and quality. In our case, the scenario was reverse: we had to work within the constraints of an organisation that already had a dataset related to procurement digitalised and a fixed problem. We share our experiences that come from several years of collaboration with A*STAR Procurement Office (A*PO), the procurement audit arm of A*STAR - a large governmental organisation in Singapore. We describe this collaboration step by step (see Sec. 3), showing how our solution evolved over time as we put emphasis on data analytics explainability. In our experience, introducing a data driven framework significantly diverting from prior procurement officers habits had a negative effect on adoption and understandability of results (see Sec. 7.2). Therefore, we took a different approach that involved procurement officers in sharing their domain knowledge to build up a set of risk scores (later referred to as risk indicators, see Sec. 5).

During the period of this collaboration we created the Procurement Analytics Core Engine (PACE) - a toolkit aimed to provide a multi-faceted and explainable solution to procurement fraud detection. By this we mean a tool that allows to investigate different kinds of fraud and capable of generating quarterly reports to the procurement officer who is fulfilling the role of an auditor. More precisely, the requirement of our collaborators was to deliver a ranking of suspicious transactions along with reasons why those transactions were selected for further investigation. This explainability of risk scores is reached mainly in two ways: (i) simplifying the risk indicators so that users can immediately understand what risk score is pertaining to; and (ii) detecting elaborate fraud schemes through ensembling simpler indicators together so that more complex risk scores can be broken down into interpretable components.

In this article we describe: the principles and formalisms driving our framework; how we made all elements come together; and how it worked in practice during system deployment. We disclose some of the details of the building blocks, however due to the confidential nature of fraud detection the presented algorithms are simplified to conceal some of the critical details yet illustrate the key assumptions of our risk indicator philosophy. Further limitations are related to the data at our disposal (see Sec. 4) - past fraud or noncompliance cases were not made available to us, therefore proposed algorithms rely mainly on unsupervised learning and statistical approaches not requiring labelled data.

The main contributions of this case study are:

1. Introduction to the procurement domain operations from a data analytics perspective and the chal-

lenges encountered in practical deployment scenarios.
2. Development of a data-driven method for procurement fraud detection using a large collection of real-world procurement data.
3. Description of experiences in successful real-life deployments of our procurement analytics software.

2. Related Work

The research on fraud detection in procurement area can be split into several approaches depending on type and quality of data used. Jans et al. (2011) notice that majority of published works on enterprise fraud are related to external fraud and mostly comprise of supervised approaches. Otherwise, what Jans et al. (2011) label as internal fraud (i.e. committed by members of organisation) is frequently investigated using unsupervised techniques due to lack of readily available past fraud cases. In comparison, we try to detect both external and internal fraud (i.e. fraud committed by either employees or vendors), however only with unsupervised means. Singh et al. (2019) like us also address both types of fraud in procurement (internal and external) but have a significant amount of past fraud cases available for analysis and therefore apply supervised learning, a technique not applicable for our case study.

Among approaches that do not require a labelled training set, there is a group of publications which try to leverage the most classical red flag-based approach (Baader and Krcmar, 2018). Velasco et al. (2021), similar as this article, present a government organisation case study, however their approach limits to detecting risk patterns without any scoring mechanism. The quantitative results in their publication show a large number of detected irregularities, however without any capability to prioritise the most critical ones for auditor. We see this as our stating point and push the state of the art further by delivering explainable scoring mechanism for detected anomalies. This goal of downsizing the samples for audit was also assumed by Moreno Oliverio et al. (2019). Their solution applies clustering to group orders characterised by features generated from simple red flags; afterwards marking cluster with highest red flag scores for investigation. In comparison to our approach, they do not point to any specific fraudulent orders or requesters but only narrow down the auditor search to a data subset. Similarly, Carlsson et al. (2018) use clustering, however theirs is a semi-supervised scenario which requires past fraud cases to identify suspicious clusters. Both Baader and Krcmar (2018) and Jans et al. (2011) take a different approach and apply process mining from ERP systems to aid fraud detection. We do not consider such approach as our goal is to maintain a working solution to fraud detection based just on a minimalist dataset. Additionally, Baader and Krcmar (2018) propose aggregating red flags under fraud patterns, albeit do it in a less formal manner than us and test only on synthetic data. In comparison to all aforementioned, Dhurandhar et al. (2015) propose a solution most similar to ours, attempting to embed more analytics into the classic red-flag approach. However, they focus on formalising concept of refining risk score through user feedback and omit problems of ensembling risk indicators and explainability brought up by us.

For fraud detection understood in a broader sense and applied to various domains there is a big number of works across the past years. As summarised by Phua et al. (2005), the most common are financial fraud and credit card fraud, however some published research can be also found in the telecommunications domain (Kaiafas, 2020) or insurance fraud (Nian et al., 2016). The proposed solutions include a wide spectrum of statistical models (Assylbekov et al., 2016), solutions based on probability theory

(Anbarasi and Dhivya, 2017) and focused on advancing technical aspects of applied machine learning (Phua et al., 2005). Majority of those pertain to supervised learning and evaluate variety of algorithms such as neural networks (Ovsyannikova and Domashova, 2020) and Bayesian networks (Arief et al., 2016), or decision trees (Correa Bahnsen et al., 2016). In comparison to our work, the key difference is the area of application. In this article, we focus on the small details that distinguish implementing fraud detection in procurement from other domains. In that sense, we see the core value of our work in aiding procurement practitioners rather than fraud research in general.

In terms of algorithmic approach, achievements in research on anomaly detection offer many different solutions applied to areas beyond fraud detection. Aggarwal (2013) reviews this entire domain and is the source of our inspiration regarding development of ensemble anomaly detection. We expand on his ideas and formalise them to work specifically for procurement application (see Sec. 3.1). Related to our work on risk indicators, the most suitable anomaly detection approaches are those not requiring labelled data, such as unsupervised learning or statistical anomaly detection (Aggarwal, 2013). However, the issue with many algorithms is the difficulty with explaining to the end user the reasons behind detected anomalies and their relationship to business context. For instance, one-class classifiers such as work of Zheng et al. (2019) adapted for fraud detection sacrifice explainability for performance. In opposition to that, our statistical outlier detection approach is tailored specifically for procurement and simplified to give better understanding of results to the end user. There are some works on combining explainability with performant AI black box models (Ribeiro et al., 2016; Lundberg and Lee, 2017), however in our case this level of insight into model results was still insufficient for the end users.

3. Organisational context: challenges and our approach

The partner agency we worked with has a procurement portal for management of all purchase orders. This portal is administered by the procurement operations team and audited by the A*PO. The latter will perform audit by processing and reviewing procurement transaction records.

During initial interviews, we noted that the main tool used by procurement officers was eyeball sampling for compliance checks and detecting fraudulent transactions. For every quarterly report auditors would pick on random a range of transactions and run a detailed investigation (sometimes subsetting based on certain criteria e.g. value). The disadvantage of such approach is intensity of human effort to perform the checks and coverage of rather limited amount of orders in comparison to the organisation total throughput of purchases. Therefore, in search of a remedy and as an experiment, the A*PO has started to explore Business Intelligence software (e.g. Tableau). This turned out to be useful for macro-level insights such as identifying departments or divisions with maximum procurement spending; or computing average expenditures. However, those tools remained limited in micro-level analysis such as detection of subtle procurement fraud (e.g. collusion between requesters and suppliers) a task previously possible with eyeball sampling albeit in limited scope.

The described workflow and operating practices in the procurement department additionally introduced a number of issues typical for manual audit and resulting from maturing corporate procedures becoming inefficient over time: tedious and error prone process; data inconsistencies and incompleteness; long turnaround time; and distributed data storage locations related to multiple input systems. Additionally, accuracy of this entire audit process was not measured in a consistent manner.

3.1. Our approach

To tackle the aforementioned challenges, answer A*POs unfulfilled needs and the growing push from Singaporean government for proliferation of data analytics in governance, we initiated a data driven approach in phases starting in 2014.

During the initial proof of concept stage we aimed to test replacement of the manual random sampling approach with an automated analysis of the entire procurement database. Our main focus was issue of poor procurement database coverage. Also, we wanted to make the entire process more quantifiable diverting some attention to measurement of accuracy. The key idea at this stage was to associate a risk score with every procurement transaction and rank the transactions based on their score. This would enable the transactions to be prioritised for manual sampling and curation.

When devising a method for computation of the risk scores, we had a number of conversations with procurement auditors. This resulted in observation that typically detected suspicious cases tend to be unusual purchases, having some abnormal activity in terms of values or purchasing patterns over time. Those characteristics would stand out from the majority of otherwise regular and sound transactions. Therefore, in terms of algorithmic approach, we decided to investigate anomaly detection methods. Narrowing down, due to lack of prior well documented fraud cases, we picked to implement an unsupervised statistical solution.

To start with, we considered three risk analyses for ranking purchase orders: 1) requester-vendor relationships; 2) potential order splits of big value into smaller, less conspicuous ones; and 3) price deviations of the same product ordered across multiple different purchase orders by different people. The core concept for each of those was to scrutinise on one fraud type separately from others; next formalise in algorithmic terms how fraud is manually identified by the procurement officers; and finally expand human approach by combining it with data analytics techniques to boost the checks in terms of scope as well as coverage (discussed further in Section 5). Preliminary tests of our algorithms showed to be successful and lead to detection of verified fraud cases. This gave the project more recognition and funding to pursue a similar approach yet in greater scale, more thought through, and with a systematic research framework. Further we describe this framework as it was constructed in the years 2014-2016 (first iteration); and subsequently 2017-2018 (second iteration) extending scope and improving the original concepts.

3.1.1. First PACE framework version (2014-2016)

For the first production version of the framework we included the 3 risk scores from our initial prototype and extended it with 7 other checks focusing on a the most basic and crucial subset of available data - analysis of purchase orders (see Appendix A).

We formalised those risk scores, now referring to them as risk indicators. Furthermore, we proposed to ensemble them sequentially or as a group, thus creating ensemble indicators capable of more complex yet traceable analysis that could be decomposed into smaller blocks for explainability (detailed in Sec. 5). This was done to answer the new key requirement that emerged from experiences with the initial prototype: procurement officers need to be able to understand the analytics output easily and comprehend why orders are pointed as suspicious. It became clear from our initial prototype evaluations that our solution would be at most only able to point suspicious transactions (i.e. candidate frauds). Eventually, procurement officers would need to take over and run a detailed investigation before any accusations or

actions could be taken.

Adding to those requirements, at this stage, the research on algorithms was further constrained by limitations of target deployment environment: our solution would be deployed immediately as being developed and continuously tested with users, therefore we focused on modular indicator architecture that could be expanded over time; secondly, reasonable execution time had to be guaranteed to rate about 20 thousand purchase orders consisting of 50 thousand order items for a single quarterly report yet considering several hundred thousand past purchases from few years back.

3.1.2. Second PACE framework version (2017-2018)

The second stage of development involved multiple lessons learned and therefore refinement of the original framework, improvement of algorithms to increase their accuracy, as well as expansion in terms of data scope by including: Invitation to Tender, Invitation to Quotation and external data with registry of companies operating in Singapore and their ownership. Additionally, we made some modifications to the output provided for the end user. One of the key lessons learned from prior stage related to impact of usability aspects on evaluation time and quality. In governmental organisations, officers are used to their specific workflow and are not particularly willing to adapt new tools. Factors like: how well they can understand algorithm suggestions and how much time it takes to follow up on a case pointed by algorithm, impacted greatly the evaluation capability and eagerness of officers to deliver feedback. Therefore, during this stage we moved away from concept of rating individual transactions to replace it with rating people and organisations. Individual order alerts created too much clutter, with clusters of similarly ranked purchases frequently pointing to the same person or supplier the principals that are suspects of committing fraud. This lesson depicts key difference of procurement fraud in comparison to other typical applications of anomaly detection in financial fraud such as credit card fraud procurement officers are more interested in big incidents involving tens or hundreds of transactions rather than individual occurrences that typically could be detected easily through conventional methods.

Following those observations, in next sections we detail the final design and architecture of our solution that emerged after both PACE framework iterations. Our process of attaining the final framework started with experiments in implementing existing state-of-the art techniques, i.e. conventional methods based on rule-engines wide spread in the industry organisations, as well as holistic end-to-end unsupervised machine learning approaches originating from anomaly detection domain (Aggarwal, 2013). In section 7.2 we explain why those were not adapted and describe other interesting side experiments. Prior to that, in next section we focus on the implemented solution and the underlying dataset description.

4. Procurement data flow and structure

The presented research was done based on data coming from A*STAR and was produced over the course of four years, 2010-2013 (inclusive). Throughout this period there were a total of 216,771 purchases recorded. In subsequent stages of our project we refreshed this dataset to include consecutive years. However, to present a consistent evaluation scenario with comparison of accuracy between different project stages, we describe experiments related only to data from the initial 4 years.

Within this dataset, the key elements that comprise a purchase are related to procedural stages of the procurement process as presented in Figure 1: (a) Invitation to Tender (ITT) or Invitation to Quotation

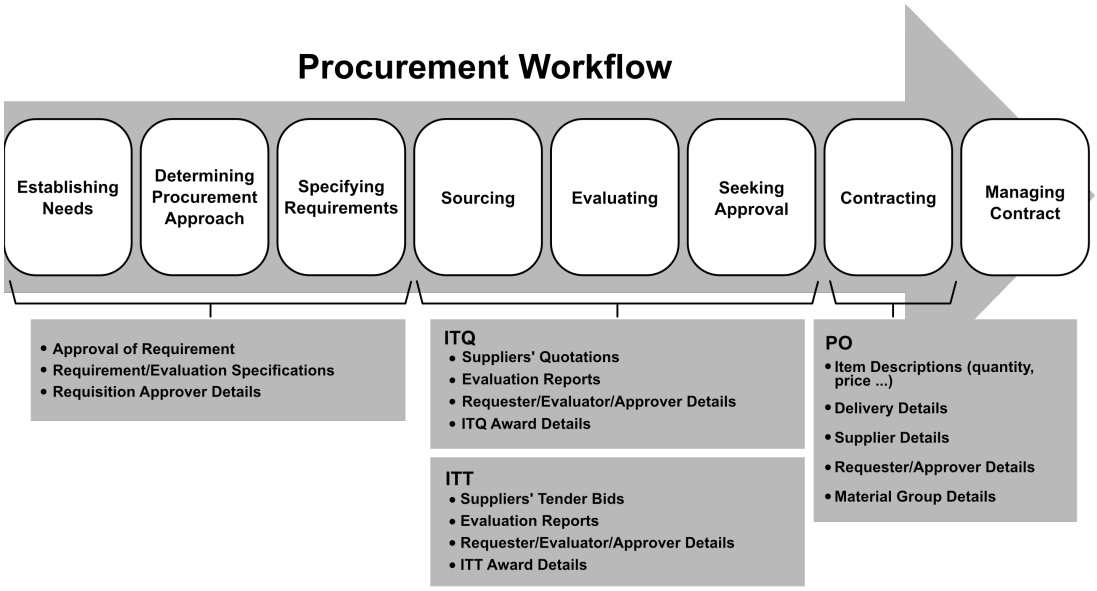

Fig. 1. Procurement workflow and its comprising steps

(ITQ); (b) bid placement and approval of selected supplier; and (c) issuance of Purchase Order (PO). The data of final stages related to delivery of goods and invoicing was not made available to us, and therefore is not included in the discussion.

The procurement process is split into multiple stages for transparency and oversight in selection of the most competitive supplier that will provide best price and conditions for the organisation. In the context of our partner, the first stage (a) could involve different level of formalities and be done in one of three following ways depending on value of purchase: 1) for very expensive purchases (tens of thousands S$ and above) tenders are required, involving a large number of formal steps and an approval committee; 2) for medium size orders (several thousand S$) buyer needs to send an invitation to quotation to three selected vendors that later come back with their offers; 3) for Small Value Purchases (SVP) a single supplier can be selected by the buyer and move directly to purchase order stage without any additional steps. From the dataset perspective the first two options have a record that describes goods to be procured along with assumed/desired price by the buyer.

For Tender and ITQ in the second stage (b) there would be a list of bids submitted by potential suppliers. Those are registered in the dataset, along with the winning bid.

Finally, when the winning bid becomes a purchase order in stage (c) it is supplied with additional information. Each of the purchase orders has a pointer to an employee of the organisation fulfilling the role of buyer (requesting officer). Similarly, each order is a subject to further approval by an approving officer. Therefore, every record has its creation date (day on which requesting officer submitted the information to the system) and an approval date. Additionally, a constraint our procurement system forces every purchase order to have only one vendor who supplies all the ordered goods of a given value expressed in local currency.

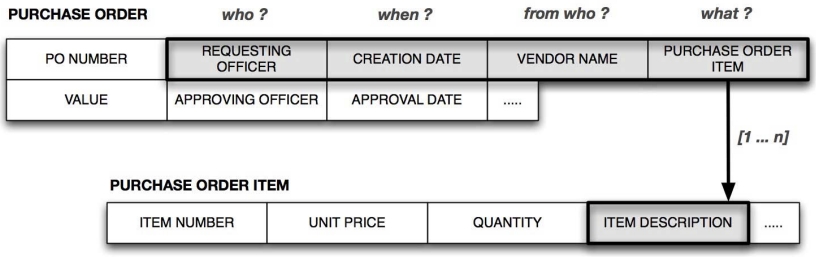

Fig. 2. Key elements of the single purchase order data vector

The purchase orders consist of purchase order items. Each of those items can relate to different goods or services and contain further details like: textual item description, quantity of items bought and unit price per single item. For example ”pantry supplies” purchase order could consist of items such as: napkins, paper cups, plastic spoons. During the earlier stages (a) and (b), there is a record with similar structure of purchases split into items with details of pricing, quantities, descriptions etc. However, the values per each stage can differ: starting with expected purchases, moving to what suppliers actually offer, ending with negotiated and approved values.

Out of all this data, during our experiments we learnt that the key elements that influenced the capabilities of our algorithms were related to quantitative relationships between the four aspects of the procurement data (see Fig. 2): who (requester), what (item description), from who (vendor), and when (creation/approval date). Requester and vendor information being particularly important as typically related to perpetrators involved in fraudulent activities. Below we detail some of the key insights that later help to understand the results obtained with our model.

The dataset being subject of our analysis contained 607,369 purchase order items. 291,333 of those records (47%) were missing requester information due to data input issues. On average, a single purchase order had attached 2.2 items but 59% of the purchase orders had only 1 order item, and 97% having 10 or less. Within the remaining 3%, the maximal recorded amount of purchase order items per purchase order was 164. This reflects the overall behaviour of the organisation employees and the policies in place that focused on simple orders, typically related to one type of good.

Table 1 contains extended statistics of the dataset, which further reveal that a significant number of requesters participated in placing orders, but at the same time, for many cases, the history of orders per requester is rather short. Together with earlier mentioned missing requester information, this posed problems for accurately applying anomaly detection on quantitative data related to requesters (e.g. requester order count, requester quarterly purchase value etc.).

Throughout our tests and deployments of PACE with different organisations, we noted that such problems tend to occur in relation to other schema nodes as well (e.g. vendors or approvers). Such data issues depended on particular organisation procedures and systemic dysfunctionalities. Since goal was to deliver a versatile solution that could take advantage of full dataset, we introduced the concept of ensemble indicators that do not rely on single part of schema. The next section details the overall framework and key constraints we put on ensemble indicators to make them resilient to data issues. Further, Section 6 presents one such ensemble indicator that heavily relies on requester data yet offsets the problem with se-

Table 1
Breakdown of key statistics for the procurement dataset as used during our experiments

Metric	Value
Purchase Order	216,771
Purchase Order Item	607,369
ITQ records	25,447
Tender	1,139
Vendor	10,556
Requester	14,641
Approval Officer	625
Unique Item Descriptions	333,228
MIN/AVG/MAX #Order PER Requester	1/ 14.8 / 1,213
MIN/AVG/MAX #Vendor PER Requester	1/ 6.3 / 187
MIN/AVG/MAX #Item PER Requester	1/ 13.5 / 979

quentially stacking analysis of other data, and another indicator that deliver results without any reliance on requester information.

5. Fraud indicator framework design

During our early experiments (see Sec. 7) and discussions with project partner we concluded that procurement fraud cannot be treated holistically as a single phenomenon. One reason is the variety of mechanisms in which fraud transpires. Another is the necessity to provide explanation to the auditors about detected cases in different ways depending on fraud type (i.e. requiring different feedback information). For example, if there is a suspicion of collusion between vendor and requester, the auditor will be most interested in all data about this relationship in comparison to other non-fraudulent relationships (frequency of interactions, value of transactions, item quantities, item categories etc.); in different scenario, when analysing concealment of true purchases through mechanism of order splits, the key data is: transaction time spacing and transaction values.

Due to this complexity, we approached every compliance check known to our collaborating procurement officers individually and formalised it separately. Each check would relate to algorithm running an analysis of procurement data dependent on nature of fraud; and subsequently output a normalised score denoting level of suspicion for a given order. Such scores would be referred to as risk indicators. At the end, all indicators would still be put into the same context of a single procurement database and come together to create a consistent image of the situation in the organisation. For all of the indicators, our intention was to establish a common denominator in form of single data schema they would feed on and a standardised quantified risk output. The underlying algorithms per indicator differ between each other quite extremely as does the nature of the fraud they are trying to detect. For example, Unusual Vendor indicator was inspired by graph analysis, while Item Spending by cluster analysis. Those indicators execute separately but afterwards our framework composes them together and creates a map for the procurement officer telling individually for every order how likely it is to be subject of particular frauds and why.

To reach this level of unification, we first formally defined the problem every indicator would have to solve regardless of fraud type; on top of it proposed the concept of indicator ensemble (as shown in next subsections); and finally assured our later design process would comply to it (see Sec. 6).

5.1. Measuring Risk: A formal indicator problem definition

Building on top of the described dataset, the central idea behind our approach is to devise a risk-based sampling methodology defined through indicators. The key problem is to find risk indicators such that optimise the identification of fraud cases. We formulate this starting from the aforementioned input data level:

Let X be a procurement data source, and x in X be a procurement transaction, where $x = x_1, x_2, ..., x_i, ..., x_k$ are the procurement data attributes. Let $X_1, X_2, ..., X_n$ be all the available procurement data sources, therefore $X_1 \, x \, X_2 \, x \, ... \, x \, X_n$ defining the joint procurement data space. Within this data space, let X_S be its data subset. A risk indicator is a [0,1] valued function f defined as:

$$f : X_S \rightarrow Y \in [0, 1] \tag{1}$$

The function f is called the base risk indicator - it can be seen that the indicator f associates a risk score in [0,1] given a procurement transaction x_i in X_S. Depending on fraud type that indicator addresses, f can be defined based on only one data source then it is implied that the other data sources are unused. Also, in other cases, f is computed based on only one transaction, which means that the subset is a singleton. For convenience, we will just use f to denote all such risk indicator functions but clarify them in context.

Ideally our goal is to obtain an indicator such that assigns a score of 1 if x_i is a fraudulent case, and a score of 0 if it is genuine. However, what constitutes fraud is ambiguous and debatable, so we modify our objective as one of finding cases that are interesting enough for further investigation. In fact, our users preferred this as it offers them flexibility to cover greater amount of suspicious cases and does not preclude the possibility of some critical fraud cases being missed out.

Therefore, we define f*, a risk indicator that assigns a score of 1 if x_i is an interesting case and a score of 0 if it is not. As a consequence, our objective is to find f such that d(f, f*) is minimised (, where d is a suitable distance function).

The challenge is that neither f* is known nor can be estimated using historical fraud cases due to their unavailability. Therefore, the task of f* approximation is handled using variety of unsupervised data analytics approaches, specific per fraud type, some of which we outlined in Section 6.

5.2. Indicator ensemble

To accommodate the indicator function for more complex scenarios, we extend its definition with the concept of indicator ensembles. The core idea here is to be able to preserve unmodified the base indicators implementing various outlier detection tasks for explainability purpose. At the same time, we would like to address more sophisticated risk functions such as multi-variate outlier detection (e.g. looking for outliers in terms of price and time constraints at the same time). This concept is motivated by the idea

of outlier ensembles earlier proposed by Aggarwal (2013). We adopt the terminology of the latter for procurement fraud risk and distinguish between two categories of risk indicator ensembles:

1. *Sequential risk* indicator ensembles, where an algorithm or a set of algorithms are applied one after another. Each subsequent indicator in the ensemble modifying the output of the prior algorithm function in execution chain:

$$f_S(x) = g(h(x)) \tag{2}$$

   where *g* and *h* are base indicators as defined previously, *x* the available procurement data source as defined in previous section, and f_S is the sequential indicator ensemble function.
2. *Independent risk* indicator ensembles, where different algorithms, or different instantiations of the same algorithm, are independently applied to either the same complete dataset or portions of the data:

$$f_I(x) = f_M(g(x), h(x)) \tag{3}$$

   where *g* and *h* are base indicator functions, and f_M is model combination function that allows to combine the independent base indicator results together. The output f_I is the independent indicator ensemble function.

An example of first case is Item Spending indicator that first employs natural language processing, followed by cluster analysis, and finally outlier detection. The example of second category is Unusual Vendor indicator, where base indicator performs outlier detection on graph to find isolated nodes with anomalous weights; and same base indicator is applied for frequency analysis to multiple data subsets denoted by various time frames. Both risk indicators have been detailed in next section.

As a further extension of the above concepts, risk ensemble functions can also be designed agnostic of fraud type and work with any kind of indicator. This allows to create tools for the auditor to adjust the overall suspicion rankings. From design perspective it also helps to further decompose indicator in smaller building blocks to improve traceability and building contextual explanations. In our final implementation of PACE 2.0 framework, we broadly used this idea for all indicators by taking the concept of a sequential risk ensemble and implementing the enveloping function as aggregate to obtain an overall scoring for requester rather than a single purchase order:

$$f_S(x) = f(x_1, x_2, x_3..., x_n) = f_A(f_D(x_i, f(x_i))) \tag{4}$$

In the above, contrary to fraud type specific risk indicators, the input is not a single transaction from X data space but multiple purchase orders. The aggregation function f_A of this indicator defines how to combine multiple purchase orders indicator values into a single score (e.g. sum, average, multiplication etc.) and can include a weight component (e.g. cumulative order value). The discriminating function f_D allows to filter out certain purchase orders based on their properties or based on their indicator functions (e.g. aggregate only purchase orders with suspicion rank greater than some threshold).

This example of implementing an ensemble indicator and the prior fraud specific ones show that our definitions of ensemble types allow some freedom when designing the final indicator. Therefore, to make the development of ensemble indicators more consistent across different implementations, we detail indicator ensemble as containing:

1. *Model components*: These are the individual methodologies or algorithms, potentially encapsulated in base indicators and integrated to create an ensemble. For example, a random subspace sampling method combines many base indicators that are applied to different subspace projections.
2. *Normalisation*: Different methods may create risk scores on very different scales. In such cases, normalisation is important for meaningfully combining the scores, so that the scores from different components are comparable.
3. *Model combination*: The combination process refers to the approach used to integrate the risk scores from different components, e.g. a weighted sum or a maximum of the individual scores.

The above properties allow ensemble functions to behave similar as regular base indicator function and therefore ensembles can be used as components in more complex ensembles. For instance, as later shown, Independent Risk Indicator Ensemble can be wrapped in Sequential Risk Indicator Ensemble. Likewise, frequently in our practice we use aggregate sequential ensembles jointly with weighting ensemble. In the following section, we describe specific examples of risk indicators and ensembles that showcase different ways of implementing such combinations in the defined framework.

6. Filling the framework with building blocks: individual indicator design

After several years of collaboration with procurement officers, the amount of fraud indicators we created has grown considerably. Our experiments outside of A*STAR have shown that indicator behaviour and usefulness are very dependent on organisation profile. Therefore, we do not give an exhaustive description of all indicators. Instead, we highlight the design process of two selected indicators that reflect well our design philosophy described in the previous section. For further reference, to gain understanding of scope and coverage of our framework, Appendix A contains an exhaustive list of all indicators along with brief descriptions. The methodology and overall design of indicators described here are reflective of what lead our project to success. However, the equations given below are only for demonstrative purposes and have been simplified in comparison to our actual deployment in order to protect the privacy of fraud detection practices in our partner institution.

6.1. Unusual Vendor indicator

The key elements that influence the fraud detection problem are related to quantitative relationships between the five aspects of the procurement data: Requesting Officer (RO), Approving Officer (AO), Vendor, Approval Date and the Total Value in local currency. Typically, the RO creates an order to purchase from a Vendor, later that order is put up to an AO for approval before the final purchase order is issued. RO can be an AO, and vice versa with the exception of same transaction. A simplified graph model depicted on Figure 3 describes this workflow.

The key observation is that graph vertices with little connections could be deemed as potentially risky and requiring more oversight. For example, requesting officer buying common goods from a vendor that nobody else deals with could indicate collusion. To reveal such cases, we note that vertices of less popular vendors will have a smaller degree in the graph than more popular vendors (e.g. Vendor Y). The unusual vendors are thus likely to be isolated from most ROs. In addition, edges of the graph

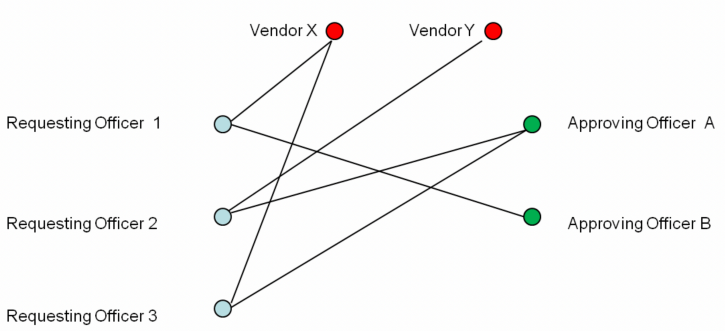

Fig. 3. Graph model describing the relationships between Requesting Officer (RO), Approving Officer (AO), Vendor

can be annotated with characteristics of the relationship, e.g. monetary value, item quantities or count of interactions in specific timeframe. If the weight of the edge significantly deviates from other edge weights for similar purchases in the context of same vendor, then the transaction suspicion can be further highlighted. Finally, splitting this analysis into timeframes enables to look at trends if the observed anomalies repeat over multiple time-frames, thus further increasing confidence to label a subgraph as anomaly.

For illustration, we demonstrate the design process following this rationale using our own dataset context: four-years of procurement transactions with about 10,556 vendors and 14,641 ROs. We modelled the interactions as a graph and analysed the degree of each vendor in the graph. A histogram of the distribution of degree values is shown in Figure 4.

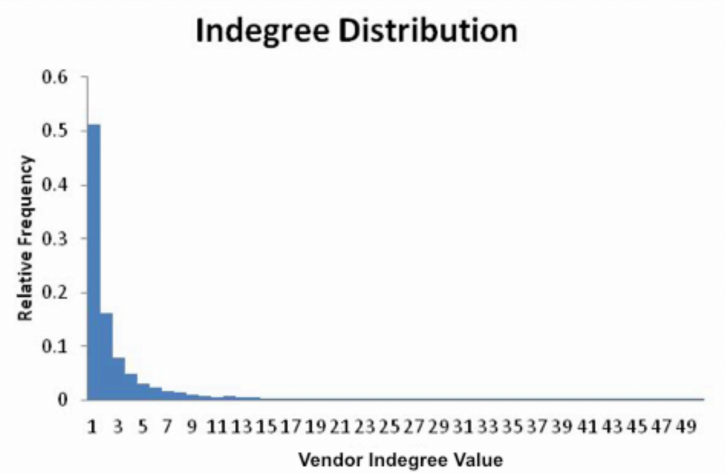

Fig. 4. Distribution of degrees in RO-Vendor graph model

It can be observed that a major portion of the vendors have small degrees, thus implying that there are a large number of vendors interacting with a small number of ROs. This means that the isolatedness

alone is not a good measure of suspicious behaviour. However, it can be useful as a starting point to identify popular vendors (in the long tail) and filter them out. To make the analysis of isolated vertices in the head of the chart finer grained, we look further into relationship analysis for count of interactions per each unique vendor-requester pair. We apply those to the graph (see Fig. 3) as edge weights and note that for individual vendors the values tend to conform to normal distribution. Therefore, we formalise this observation as our first base indicator using z-score for detecting pairs with unusual amount of interactions:

$$UVI_{CNT}(x) = \frac{avg_k(po_{cnt}(vend(x), req(k)) - po_{cnt}(vend(x), req(x))}{SD_k(po_{cnt}(vend(x), req(k))) * DIST_{MAX}(x)} \tag{5}$$

In the above, *vend(x)* is vendor involved in PO "x"; *req(x)* is the requester involved in PO "x"; po_{cnt}*(vend(x), req(x))* is amount of POs made between vendor and requester; $avg_k(po_{cnt}$*(vend(x), req(k)))* is the average count of POs across k requesters (where k are all requesters in analysed dataset X_S). Finally, in the denominator: SD_k is standard deviation of count of POs across k requesters; the $DIST_{MAX}$*(x)* is a component used for normalisation and is the maximum possible SD distance from average calculated for vendor(x). The intuition behind this final component is to calculate the maximum possible value of z-score so that UVI_{CNT}*(x)* meets the formal indicator definition from Section 5.1. In our context, this maximum value occurs when all requesters but one have the same order count. Therefore, the distance value is independent from actual count of POs between each pair *requester(x)* with *vendor(x)*, and only dependent on total requester count related to *vendor(x)*. $DIST_{MAX}$ calculation can be simplified assuming the case of all requesters having 1 PO and one having 2 POs:

$$DIST_{MAX}(x) = abs(\frac{AVG_{MAX}(x) - 2}{SD_{MAX}(x)}) \tag{6}$$

$$AVG_{MAX}(x) = \frac{((req_{cnt}(x) - 1) * 1 + 1 * 2)}{req_{cnt}(x)} \tag{7}$$

$$SD_{MAX}(x) = \sqrt{VAR_{MAX}(x)} \tag{8}$$

$$VAR_{MAX}(x) = \frac{(req_{cnt}(x) - 1) * (AVG_{MAX}(x) - 1)^2 + 1 * (AVG_{MAX}(x) - 2)^2}{req_{cnt}(x)} \tag{9}$$

AVG_{MAX}*(x)* is the average order count for *vendor(x)*; SD_{MAX}*(x)* is standard deviation of order count for *vendor(x)*; VAR_{MAX}(x) is variance of order count for vendor(x), all those values calculated for situation when all requesters but one have the same order count.

Calculated in such way, the base indicator reflects the earlier observations regarding vendor transaction frequency, adds a new dimension to analysis (i.e. requester interaction count) and holds the normalisation component earlier described in formal definition (see Sec. 5). As a result, it significantly narrows down

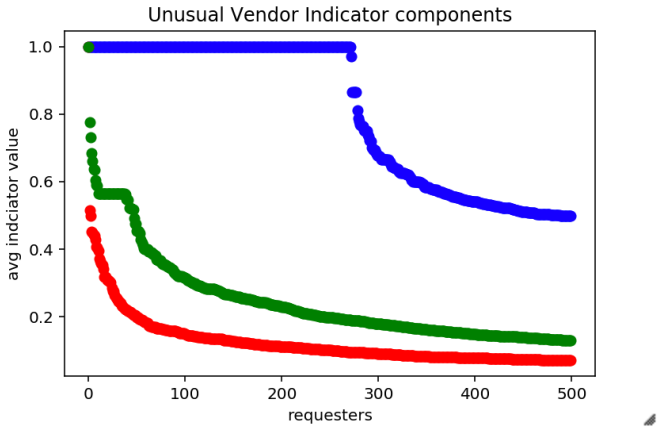

Fig. 5. Distribution of requester indicator values for the top 500 cases: (blue) UVI$_{CNT}$ order count base indicator; (red) UVI$_{VAL}$ order value base indicator; (green) UVI(x) - final ensemble

the highlighted cases in our scenario by an order of magnitude (i.e. from thousands of suspicious cases to hundreds). Following a manual evaluation with our procurement office auditors, we decided to attach one more dimension - total transaction value. We formalised it in a similar manner as transaction count but used as second base indicator:

$$UVI_{VAL}(x) = \frac{avg_k(po_{val}(vend(x), req(k)) - po_{val}(vend(x), req(x))}{SD_k(po_{val}(vend(x), req(k))) * DIST_{MAX}(x)} \tag{10}$$

The interpretation of this indicator equation is similar as previous just applied to order value rather than count. In order to combine both analyses, we formulate an indicator ensemble using above as base indicators and apply them over multiple time frames (in our case, applied yearly). We use additional weights when combining the indicators to obtain average score:

$$UVI(x) = \sum^{yearly\ periods} (w_1 * UVI_{cnt}(x) * w_2 * UVI_{val}(x)) \tag{11}$$

The final result in comparison to the intermediate base indicators can be seen on Figure 5. It can be observed how the introduction of new ensembles helped to narrow down the search for outliers. Also, due to such component design, we can easily decompose this final ensemble score and trace back to individual base indicators showing the end user if high suspicion is related to vendor isolation, order count or order pricing (or potentially all). Furthermore, since both base indicators are based on assessment of order relationship properties, we can easily link the related orders further contextualising the case by case explanation.

In practice, when applying this indicator for end users, adding base indicators to the ensemble was an iterative process, working with the users to meet their expectations and listening to experienced auditor

feedback. Additionally, during the study we added some further limitations related to business regulations: threshold for the total order value was set at S$100,000 to eliminate very big purchases that are already heavily scrutinised; the minimum number of RO connections with a vendor in a subgraph was set to 5. This entire methodology and the core concept of Unusual Vendor detection contributed to the initial success of our first PACE proof-of-concept prototype.

6.2. Item Spending indicator

The second indicator example shows a different type of analysis to reveal more clearly the particular PACE framework design choices we discussed earlier: rather than Independent Risk Ensemble it demonstrates the Sequential Risk Ensemble technique.

Instead of looking at relationship pattern abnormalities discussed in previous section, for this indicator we interpret anomalies as numeric values deviating from reference levels charted by trends. In comparison to UVI_{VAL} we put more emphasis on specific items rather than total values. The core rationale for Item Spending Indicator from fraud perspective is finding suspicious purchases through detection of items that are bought at elevated prices or non-market prices, in odd quantities or frequencies. This relates to type of fraud where: requester repeatedly colludes with vendor making purchases that are never delivered; requester manufactures factitious vendors to setup purchases that never happen; or requester obtains actual goods but at elevated pricing.

In the context of earlier presented formal indicator framework, the function implemented by Item Spending indicator has to account for unit pricing of all items participating in PO:

$$ISI_{SCR}(x) = \frac{\frac{\sum_{i=1}^{item\_count(x)} avg_{unit\_price} - unit\_price(x_i)}{sd_{unit\_price}}}{item\_count(x)} \tag{12}$$

In the above equation *item_count(x)* is count of different items bought as part the PO x; *unit_price(x_i)* is the unit price of item x_i belonging to PO x; avg_{unit_price} and sd_{unit_price} are the average and standard deviation of unit prices across all PO in the database holding similar item descriptions.

The key difficulty of the above is to find the best way of identifying item similarity. It is not a trivial task due to noisy data that is often the issue of procurement systems where many stakeholders input the data, e.g.: procurement officers, representatives of departments in an organisation, regular employees requesting certain orders etc. As mentioned earlier (see Sec. 4), in our case the entire item description was encapsulated in one manually input string. As a result we observed the following outcomes:

- Orders of different types are often placed in the system under similar or exactly the same natural language labels causing regular anomaly detection solutions to red-flag excess amount of cases.
- Orders which refer to exactly the same items are labeled differently and as a result anomaly detection algorithms omit them during analysis. In extreme cases this results in little data to analyse because seemingly there are no repeated purchases at all.

The solution proposed by us involves two steps: (1) mash the raw data looking at text through various natural language processing techniques; (2) create a structured data aggregation layer between the

Fig. 6. Data subset extracted using keyword index with reference to macbook keyword. The chart presents distribution of item unit prices on Y axis (in SGD) and consecutive orders on X axis. The red line indicates average unit price level.

anomaly detection algorithm and the original dataset. The first step helps to break up the dataset into finer grains through looking at item description in more detail. We identify each order by more than just a single string to aggregate together purchases based on new description (e.g. frequent keywords). The purpose is to remedy the aforementioned problem of multiple slightly different descriptions pertaining to same item purchases. However, we take note this step is bound to create more data complexity and consequently computational problems for the anomaly detection. Furthermore, now the more fine-grained order descriptions can result in aggregating non-related purchases into the same bucket. This problem is visualised in Figure 6. To solve the newly created problem, in the second step, we propose to analyse meaningful structured data features such as unit prices of purchased items and apply clustering before the anomaly detection algorithm. Afterwards, run the anomaly detection individually per each cluster.

In our particular deployment, for the first step we tested three different methods: (1) exact item description matching as baseline solution; (2) detect similar item descriptions using Levenstein distance measure; and (3) frequent keyword indexing. For the second step we applied k-means clustering algorithm on item unit prices. The anomaly detection layer was based on statistical deviation from the norm. A visualisation of this method using deviation from average as anomaly measure can been seen on Figure 7.

The described algorithm steps can be formalised as sequential ensemble indicator with earlier mentioned ISI_{SCORE} delivering final scoring and prior functions altering the input space as presented on Figures 6 and 7:

$$ISI(x) = ISI_{SCR}(ISI_{CLS}(ISI_{NLP}(x))) \tag{13}$$

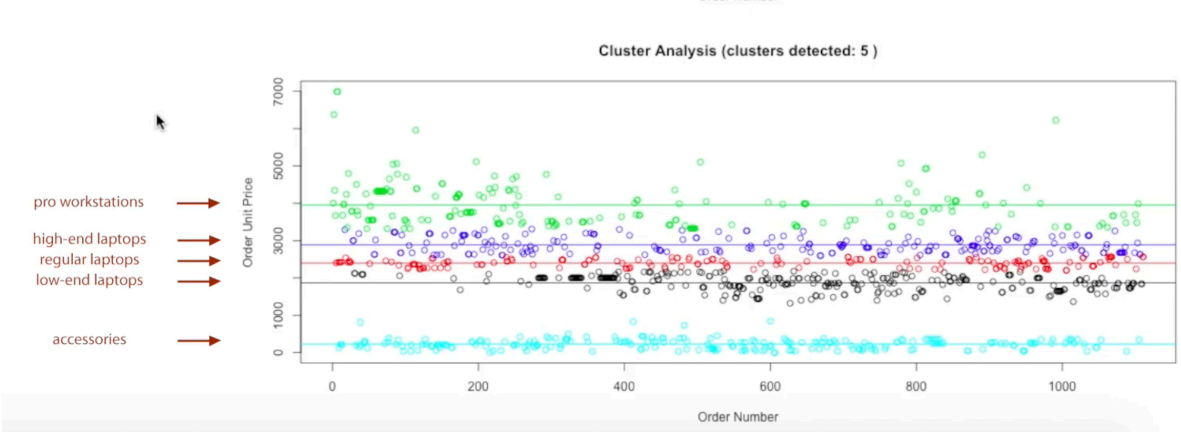

Fig. 7. Clustering over data subset extracted using keyword index with reference to macbook keyword. The chart presents distribution of item unit prices on Y axis and consecutive orders on X axis.

$$ISI_{NLP} : X \rightarrow Z; ISI_{CLS} : Z \rightarrow W; ISI_{SCR} : W \rightarrow U \tag{14}$$

$ISI_{NLP(x)}$ is the function that maps original purchase order space X into space Z consisting of POs associated with keywords. Function $ISI_{CLS}(z)$ further transforms this space Z into space W consisting of purchase orders attached to cluster numbers. Finally, earlier defined ISI_{SCR} works as final outlier scoring function delivering ranks for purchase orders.

7. Framework evaluation

Following the design of indicators as shown in previous sections, we evaluated them in stages accordingly to version of framework, i.e., PACE 0.5 evaluation, 1.0 evaluation and 2.0 evaluation.

The first prototype (PACE 0.5) involved a simple proof-of-concept model and therefore its evaluation was done informally. Firstly, we produced 3 separate rankings for each of the 3 implemented indicators. Each ranking contained orders from the entire procurement database. Afterwards, we delivered those results to the procurement officers and allowed them to investigate freely. The officers evaluated about 20-30 purchase orders per indicator. Intuitively, we expected Item Spending to give best results as this indicator was designed following the most typical unit price and purchase value checks performed earlier by officers manually. In practice however, the feedback was for Unusual Vendor delivering the most interesting results, followed by Order Split. In particular, Unusual Vendor helped to detect multiple fraud cases related to a single requesting officer. Order Split did not result in any serious frauds but depicted multiple suspicious cases that pointed to lapses resulting from incompliance or exploiting procurement procedures. For Item Spending, evaluators complained on results being too obvious and top scores allocated to purchases that were deviating from norm for expected reasons (e.g. rare big term contracts or bulk purchases at preferable prices thus deviating for their historical values). In addition, the evaluation

Table 2
PACE 1.0 Evaluation results (Overall score is the average across all indicators)

No	Indicator Name	Precision@20
1	Split order	6%
2	Digit Frequency	50%
3	Border Value	30%
4	Duplicate Order	30%
5	Unusual Vendor	50%
6	Name Patterns	46%
7	Round Value	40%
8	Requestor Spending	34%
9	Item Spending	15%
10	Vendor Spending	15%
	OVERALL	30%

showed issues with quality of textual item descriptions that played key role for this indicator. It turned out multiple different items were holding same or very similar descriptions due to non-standardised manual input done by hundreds of different requesting officers. We attempted to address those issues in the next version of PACE framework (using Item Spending implementation as discussed earlier in Sec. 6.2).

The second iteration of framework (PACE 1.0), took a more formal approach to evaluation following a more elaborate research and engineering processes. At this stage we included 10 indicators based on analytics for Purchase Order dataset. For each indicator we took top 20 recommendations as ordered by indicator score and passed to procurement officers for evaluation. For each of the indicators we measured precision@n as follows:

$$precision@n = \frac{orders\ \ confirmed\ \ suspicious\ \ by\ \ auditor}{n} \tag{15}$$

In the above equation, orders confirmed as suspicious are of a subset of cases recommended to the auditor which were subsequently identified as possible frauds and worthy of further investigation. Using this definition, the average precision@20 across all 10 indicators for PACE 1.0 gave result of 31.6%. The detailed results per each indicator can be seen in Table 2.

The final framework version (PACE 2.0) was evaluated in a similar way as PACE 1.0, however: a) number of evaluated indicators was greatly extended (including updated versions of PACE 1.0 indicators); b) majority of indicators were evaluated several times as they were refined with new suggestions from procurement officers (2-4 times depending on indicator and evaluation results). The results for the same subset indicators as in PACE 1.0 improved significantly delivering average 76% precision@20, up from 31.6%. For the full set of indicators including newly developed ones, we achieved p@20 equal to 67.1%. The detailed results can be seen in Table 3.

Table 3
Results of PACE 2.0 evaluation. Summary of performance for different indicator categories (overall score is the average across all indicators of all categories)

No	Indicator Name	Precision@20
1	PO	76.0%
2	ITT	26.0%
3	ITQ	72.5%
4	ACRA	72.3%
5	HYBRID	71.6%
	OVERALL	67.1%

7.1. HCI issues and their impact on evaluation

Throughout the project duration, PACE framework output and its presentation to end user constantly evolved. This happened due to difficulties related to evaluation of fraud detection in procurement, i.e. often the procurement officers were unable to tell if an order suggested by our algorithm is indeed suspicious or not in reasonable time and with certainty.

During the initial stages of our project, it was common for procurement officers not to understand why algorithm highlighted an order as suspicious. As a result many algorithm recommendations were marked as false positives (in order to speed up the evaluation task). Therefore, after conversations with our evaluators we modified the algorithms and the way results were displayed.

PACE 0.5 presented users with a separate page per each of the three implemented indicators. Each page contained a ranked list of orders - the rank being an unnormalised number equal to direct output a particular indicator calculation (e.g. z-score for price deviations). Additionally, for every indicator we provided some basic purchase order information (e.g. requester, vendor, item description, value, PO number).

For PACE 1.0, we experimented with two new views for different kinds of investigations: 1) ranking of purchases based on order suspicion rank - an ensemble indicator deriving from all indicators implemented at the time in the framework. In this case, auditor could also see individual indicator ranks contributing to the final score and navigate to related orders that impacted each base indicator calculation. This allowed to obtain a partial explanation and hints what to look at during investigation (see Fig. 8); and 2) second view, offered a ranking focused only on single base indicator score similar as in PACE 0.5. However, this time we also presented related orders that contributed to the score. Furthermore, apart of UI changes, all indicators and their related equations were explained to the evaluators for deeper understanding of results.

Finally, in PACE 2.0, we deviated from the concept of order ranking and replaced it with requester, approver and vendor rankings. This made each position in the ranking list a cluster of suspicious purchases rather than a single order as before. Additionally, some weighting mechanisms previously embedded into the indicator algorithms were pulled out and provided as options for every indicator. For example, regardless of indicator type, auditor could impose additional weights based on order value, thus promoting or demoting big purchases.

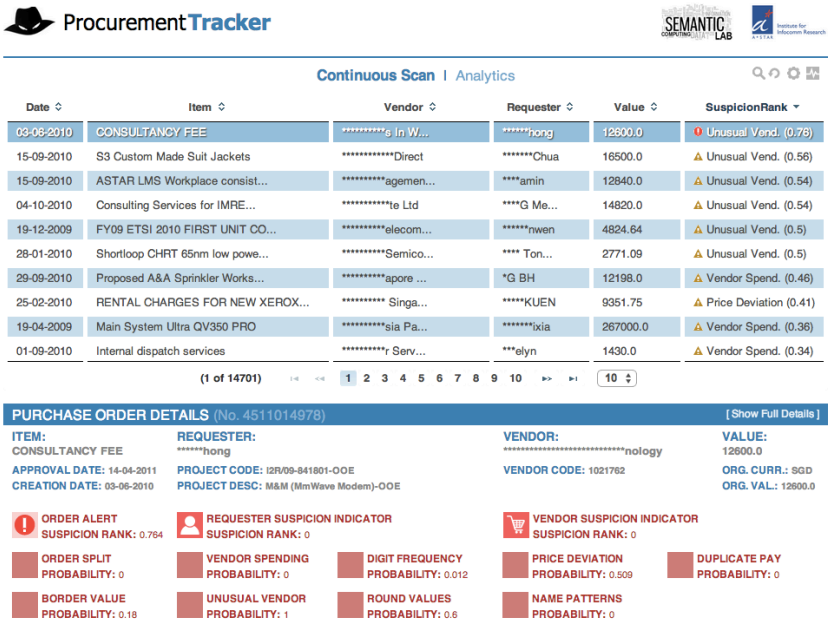

Fig. 8. PACE 1.0 showing ranking of orders based on aggregate score.

7.2. Evaluation of alternative approaches

Aside of the described evaluation and the developed PACE framework, we did some additional experiments to check the performance of other alternative approaches used by practitioners and proposed in academic literature.

During a survey of other governmental institutions, we noted that usage of rule engines is a popular approach for compliance checks and fraud detection. The key mechanics in those engines relates to simple binary checks executed for all orders and verifying presence of values in certain columns in the database; or fulfilment of certain conditions by a combinations of columns. The advantage of this approach is its simplicity and immediate actionability for the auditor without much further investigation. On the downside, the simplicity of logic results in institutions maintaining huge amount of such rules to reflect the business complexity. This leads to many orders highlighted as suspicious and officers needing to resort again to random sampling albeit from a smaller pool than previously with traditional audit. As our partner and project sponsor were not interested in taking this direction, we didnt evaluate the accuracy of such solution in the scenario of our dataset.

On the opposite side of the spectrum, we investigated the academic approach to fraud detection based on unsupervised machine learning. We picked a popular algorithm adapted to anomaly detection suggested by literature to perform well in fraud detection cases and tested it on our dataset. Namely, we used k-NN (Ramaswamy et al., 2000) attempting for similar evaluation as with our indicator framework. We took top 20 ranked cases by k-NN and passed them to evaluators to confirm if the orders were suspicious or not. However, this time such evaluation approach turned very problematic - the procurement officers

would not know what to look for and always marked every order as not suspicious. We tried to resolve this issue by using some explainable AI approaches (Molnar, 2019) that help to point out features contributing to algorithm judgement. However, this was still not enough and our results did not improve. Subsequently, unable to deliver a credible evaluation and useful results for a practical deployment we abandoned this direction in favour of a middle approach between the rule engine and more elaborate machine learning - our PACE framework.

8. Conclusions and Future Work

Technologies that harness Data Analytics can be used to obtain actionable insights from volumes of data captured during the daily procurement operational activities. However, our experiences show that those technologies have to be used carefully and well thought through when deployed in organisation as part of new processes in order to bring expected benefits. Our experiments with end-to-end machine learning show that in practical scenarios we have to be cautious to resist the managerial push to adopt cutting edge technologies and new hot trends. Furthermore, in the procurement fraud domain, we have to note that same tools and algorithms have very different results depending on organisation profile. Apart of the described experiments in the evaluation section, we did a series of trials for PACE framework in multiple other organisations operating in Singapore. This includes similar government agencies, ministries, other public organisations and private national/ multinational enterprises. Aside of experiments described in this article, we did 10 other methodologically similar, single round evaluations that resulted in p@20 spanning from 0.2 up to 0.8 depending on organisation. The reasons for such a big range of results were multitude but most often connected different procedures in organisations. Certain practices considered as gaps exploited by requesters, would not exist in other organisations due to management differences. In even more extreme cases, forbidden or incompliant actions in some organisations, would be turned into actual operating procedures in others. For instance, in some organisations splitting purchases would be considered incompliance. In others, this was part of standard operating procedures for certain orders or goods. Furthermore, although many organisations share similar data schema for procurement, we noted that due to differences in management and internal procedures, the way this schema is populated with actual data can change the meaning of what is analysed. For instance, the Unusual Vendor indicator described earlier helped us to detect fraud in one of the organisations, while in other was a reason for multiple false positives due to having designated person from organisation doing exactly such isolated purchases on behalf of entire company staff.

Those and other particularities of implementing procurement processes in organisation caused our engine to behave with varying accuracy. Therefore, in this article we did not focus on details for individual indicator algorithms but the overarching approach which we see as the main contribution. We advance the prior state of the art with a middle ground approach that delivers ranking capabilities similar to unsupervised learning approaches but explainability more akin to rule-engines. We expand of the previous definitions of outlier ensembles and make it practical for procurement fraud application.

Aside of fraud detection our collaborators from A*PO naturally saw the PACE platform as tool for compliance checks and helped us to design some of the indicators using their internal audit procedures. At the time of development of the framework, the office would go through a set of 68 compliance checks per each audited order. We implemented 7 out of those as part of PACE indicators (remainder were

impossible due to not all data digitised). As calculated by A*PO, this resulted in time/cost savings of about 10%.

This and other mentioned benefits contributed to the successful deployment of the framework. However, moving forward we still see limitations and potential for growth. In our view among the key challenges we face is a solution for efficient and accurate evaluation that would allow to tune platforms such as ours and others mentioned in related work. Adjusting the weights with user feedback is a problem already approached by Dhurandhar et al. (2015), however we find it still unsolved in practical deployments with limited number of auditors. Due to such problems, we see a more thorough evaluation of ensemble indicators and further development of explainable fraud detection as subjects of future work.

Acknowledgments

We would like to express our gratitude to Samuel Ong and Henry Chang, formally from A*STAR Procurement Office, for their help and input during various stages of our research.

References

Aggarwal, C., 2013. *Outlier Analysis*. Springer, New York.

Anbarasi, M.S., Dhivya, S., 2017. Fraud detection using outlier predictor in health insurance data. In *International Conference on Information Communication and Embedded Systems (ICICES)*, pp. 1–6.

Arief, H.A., Saptawati, G.A.P., Asnar, Y.D.W., 2016. Fraud detection based-on data mining on indonesian e-procurement system (spse). In *International Conference on Data and Software Engineering (ICoDSE)*, pp. 1–6.

Assylbekov, Z., Melnykov, I., Bekishev, R., Baltabayeva, A., Bissengaliyeva, D., Mamlin, E., 2016. Detecting value-added tax evasion bybusiness entities of kazakhstan. In *Intelligent Decision Technologies 2016*, pp. 37–49.

Baader, G., Krcmar, H., 2018. Reducing false positives in fraud detection: Combining the red flag approach with process mining. *International Journal of Accounting Information Systems* 31, 116.

Bartolini, A., 2011. CPO 2011: Innovative Ideas for the Decade. Technical report, Ardent Partners Research.

Buffett, S., Scott, N., 2004. An algorithm for procurement in supply-chain management. In *AAMAS 2004 Workshop on Trading Agent Design and Analysis*.

Calafato, A., Colombo, C., Pace, G.J., 2014. A domain specific property language for fraud detection to support agile specification development. In *CSAW 2014 Computer Science Annual Workshop Malta*.

Carlsson, C., M., H., Wang, X., 2018. Fuzzy c-means for fraud detection in large transaction data sets. In *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pp. 1–6.

Correa Bahnsen, A., Aouada, D., Stojanovic, A., Ottersten, B., 2016. Feature engineering strategies for credit card fraud detection. *Expert Systems with Applications* 51, 134–142.

Dhurandhar, A., Ravi, R.K., Graves, B., Maniachari, G., Ettl, M., 2015. Robust system for identifying procurement fraud. In *Proceedings of the Twenty-Seventh Conference on Innovative Applications of Artificial Intelligence*, p. 38963903.

Grand View Research, 2019. Procurement As A Service Market Size, Share & Trends Analysis Report By Component (Strategic Sourcing, Category Management, Contract Management), By Organization, By Vertical, By Region, And Segment Forecasts, 2019 2025. Technical report.

Jans, M., Lybaert, N., Vanhoof, K., 2010. Internal fraud risk reduction results of a data mining case study. *International Journal of Accounting Information Systems* 11, 1, 17–41.

Jans, M., M., v.J., N., L., K., V., 2011. A business process mining application for internal transaction fraud mitigation. *Expert Systems with Applications* 38, 1335113359.

Kaiafas, G., 2020. Ensemble Learning for Anomaly Detection with applications for Cybersecurity and Telecommunications. Ph.D. thesis, University of Luxembourg.

Lundberg, S.M., Lee, S., 2017. A unified approach to interpreting model predictions. In *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, p. 47684777.

Molnar, C., 2019. Interpretable machine learning. a guide for making black box models explainable. https://christophm.github.io/interpretable-ml-book/.

Moreno Oliverio, W.F., Silva, A.B., Rigo, S.J., Bezerra da Costa, R.L., 2019. A hybrid model for fraud detection on purchase orders. In *Intelligent Data Engineering and Automated Learning IDEAL 2019. Lecture Notes in Computer Science, vol 11871*.

Mori, M., Kobayashi, R., Samejima, M., Komoda, N., 2013. An evaluation method of reduced procurement risks by decentralized ordering in supply chain. In *11th IEEE International Conference on Industrial Informatics (INDIN)*.

Nian, K., Zhang, H., Tayal, A., Coleman, T., Li, Y., 2016. Auto insurance fraud detection using unsupervised spectral ranking for anomaly. *The Journal of Finance and Data Science* 2, 1, 58–75.

Ovsyannikova, A., Domashova, J., 2020. Identification of public procurement contracts with a high risk of non-performance based on neural networks. *Procedia Computer Science* 169, 795–799.

Phua, C., Lee, V., Smith, K., Gayler, 2005. A Comprehensive Survey of Data Mining-based Fraud Detection Research. Technical report, Clayton School of Information Technology, Monash University.

PwC, 2014. PwC Global Economic Crime Survey 2014. Technical report.

Ramaswamy, S., Rastogi, R., Shim, K., 2000. Efficient algorithms for mining outliers from large data sets. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, p. 427438.

Ribeiro, M.T., Singh, S., Guestrin, C., 2016. "why should i trust you?": Explaining the predictions of any classifier. In *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, p. 11351144.

Ruping, S., Punko, N., Gunter, B., Grosskreutz, H., 2008. Procurement fraud discovery using similarity measure learning. *Tran. CBR* 1, 1, 37–46.

Singh, N., Lai, K., Vejvar, M., Cheng, T.C.E., 2019. Data-driven auditing: A predictive modeling approach to fraud detection and classification. *Corporate Accounting & Finance* 30, 64–82.

Supply Management. Chartered Institute of Procurement & Supply (CIPS), 2014. Procurement fraud second most common economic crime globally. http://www.supplymanagement.com/news/2014/procurement-fraud-second-most-common-economic-crime-globally.

Umbenhauer, B., Gregson, J., 2016. The Deloitte Global CPO Survey 2016. Technical report, Deloitte.

Velasco, R., Carpanese, I., Interian, R., Paulo Neto, O., Ribeiro, C., 2021. A decision support system for fraud detection in public procurement. *International Transactions in Operational Research* 28, 27–47.

Wang, G., Miller, S., 2005. Intelligent aggregation of purchase orders in e-procurement. In *Proceedings of the 2005 Ninth IEEE International EDOC Enterprise Computing Conference (EDOC05)*.

Yuan, H., Sun, X., Jing, Y., Pan, W., Ren, T., 2011. Design of material ordering strategy in procurement supply chain. In *Chinese Control and Decision Conference (CCDC)*.

Zheng, P., Yuan, S., Wu, X., Li, J., Lu, A., 2019. One-class adversarial nets for fraud detection. pp. 1286–1293.

Appendix A

Table A1
List of fraud indicators.

Type	Indicator Name	Scope*	Description
non-compliance	Duplicate Pay	PO	Order was duplicated multiple times during the same day
non-compliance	Order Split	PO	Order is involved in scheme by requester splitting bigger purchases into smaller ones
fraud	Item Spending	PO	Unit prices of items within a PO far from the average for those items as found in past orders
fraud	Vendor Spending	PO	PO value differs significantly from the average order value for the particular vendor
fraud	Border Value	PO	Order value is close to the limit for un-tendered order
fraud	Round Value	PO	Order value is a round number
fraud	Name Patterns	PO	Order has similar vendor name to other orders made by the same requester
fraud	Unusual Vendor	PO	Order was made with a vendor that has little interactions with any other requesters except of one
fraud	Benford Analysis	PO	Order belongs to a subset of purchases where value digit frequency deviates from the reference frequency of Benford's Law
fraud	Requester Spending	PO	PO value differs significantly from the average order value for the particular requester
fraud	Shell Companies	ACRA	Requester buying from different suppliers, with different directors but both present in some other 3rd supplier
fraud	Common Directors	ACRA	Requester making purchase from 2 different suppliers that share at least 1 director
non-compliance	Conflict of Interest	ACRA	Requester making purchase from a company where he is also a director
non-compliance	Quick Close	ITT	Time difference between tender notice and closing dates significantly smaller in comparison to typical cycles for past tenders
fraud	Fast Evaluation	ITT	Time difference between tender closing and final internal evaluation dates suspiciously small in comparison to past tenders
fraud	Closest Winner	ITT	Winning tender bid and the next closest bid are very close to each other in terms of value
fraud	Overpriced Award	ITT	Winning bid is not the best cost-wise option in comparison to other bids submitted
fraud	Award Similarity	ITT	Estimated tender value by requester is very close to actual awarded bid
fraud	Bid Similarity	ITT	Estimated tender value by requester is very close to submitted bid values
fraud	Border Value	ITT	Estimated tender value is close to the pre-defined border (for example tender value that involves additional approvals)
fraud	Benford Analysis	ITT	Estimated tender value is in a subset of purchases with digit frequency deviating from the reference of Benford's Law
fraud	Round Values	ITT	Total awarded value for tender is a round number
fraud	Sensitive Procedure	ITT	Sensitive tender procedure used to manage tenders of a requester
fraud	Sensitive Category	ITT	Tender related to purchase of specific type of goods (possible more prone to fraud than others)
fraud	Unusual Vendor	ITT	Tender was awarded to a vendor that has little interactions with any other requesters except of one
fraud	Frequent Invite	ITT	Vendor being frequently invited in tenders by selected requester
fraud	Frequent Award	ITT	Vendor being frequently awarded in tenders by selected requester
fraud	Single Bid	ITT	Requester participating in tenders that have only a single bidder
fraud	Name Patterns	ITT	Tender has similar awarded vendor name to other tenders made by the same requester
non-compliance	Tender Split	ITT	Tender is involved in a scheme by requester splitting bigger purchases into smaller ones
fraud	Frequent Invite	ITQ	Requester frequently inviting the same supplier
non-compliance	Absurd Estimates	ITQ	Proportional to ratio of APV/EPV (APV = Awarded Purchase Value, EPV = Estimated Purchase Value)
fraud	Fabricated Vendor	ITQ	Requester invites a vendor that is never bidding (ie. fabricated competitor)
fraud	Border Value	ITQ	EPV (estimated purchase value) is close to the 70k border (tender border)
fraud	Award Similarity	ITQ	EPV is very close to APV value, the closer the more suspicious
fraud	Frequent Award	ITQ	Requester frequently awards the same supplier
fraud	Single Bid	ITQ	Requester frequently getting a single bid (more suspicious if happens more frequently)
fraud	Sensitive Procedure	ITQ	Frequency of use of limited ITQ by given requester (the more frequently the more suspicious)
fraud	Frequent Spender	ITQ	Frequency of purchases (the theory is that frequent purchasers could be suspicious)
non-compliance	Quick Close	ITQ	Time difference between ITQ notice and closing dates significantly smaller in comparison to typical cycles for past ITQs
fraud	Fast Approval	ITQ	Time difference between ITQ creation and approval dates suspiciously small in comparison to past ITQs
fraud	Late Awarded Bidder	ITQ	Same vendor always coming as last bidder and winner (collusion with requester to get other bid data)
fraud	Frequent Looser	ITQ	Supplier frequently not awarded
fraud	Lucky Winner	ITQ	PO with requester-vendor was not awarded but afterwards requester did another EPR and awarded same vendor
non-compliance	Excess Spending	HYBRID	PO value (PO dataset) greater than APV (ITQ dataset)
non-compliance	Unawarded Vendor	HYBRID	Different supplier awarded (ITQ dataset) and different supplier for PO (PO dataset)
non-compliance	Early Approval	HYBRID	PO approved date (PO dataset) is earlier than EPR approved date (ITQ dataset)
non-compliance	Ghost Vendor (ACRA)	HYBRID	Awarded supplier is not active at the time of award

*Scope refers to part of dataset used by indicator: PO (Purchase Order), ACRA (registry of companies with directorship etc.), ITT (Invitation to Tender), ITQ (Invitation to Quotation), HYBRID (using multiple datasets from the aforementioned ones)