Vol. 4, No. 1, pp.: 57-70, 2026

ProcureAnomalyDB: A Public Procurement Database for Fraud, Collusion, and Market Concentration Analysis

Rui Tavares Almeida[1], Beatriz Carvalho Pinto[2, *], João Magalhães Ribeiro[3], Inês Marques Silva[4]

[1] Department of Information Systems, University of Trás-os-Montes and Alto Douro, Vila Real 5000-801, Portugal

[2] Department of Public Administration, University of Évora, Évora 7004-516, Portugal

[3] School of Technology and Management, Polytechnic Institute of Leiria, Leiria 2411-901, Portugal

[4] Department of Economics, University of the Azores, Ponta Delgada 9500-321, Portugal

** bcpinto@uevora.pt*

Article Information	
Received	17 October 2025
Accepted	28 February 2026
DOI	https://doi.org/10.63646/datamind.2026.040105

Abstract

Public procurement accounts for between twelve and twenty percent of gross domestic product in OECD member states, and the published microdata describing tenders, bidders, bids, awards, and complaints has become one of the largest and most consistently updated administrative datasets that competition authorities, audit offices, and supreme audit institutions can access. Yet despite this volume, integrity oversight is still typically conducted through manual sampling and ad-hoc spreadsheets, because no widely adopted research database treats the six entity classes of public procurement as a coherent system. This article presents ProcureAnomalyDB, a public procurement database whose schema, field dictionary, indexes, quality-control pipeline, ethics regime, and reusable application programming interface are organized around three integrity questions: which tenders show signs of collusive bidding among co-bidding cohorts, which bids are anomalous in value relative to expected price, and which buyer-supplier pairings exhibit unusual market concentration. Six core entities (NOTICE, BIDDER, BID, AWARD, COMPLAINT, ANOMALY_LABEL) are organized so that every flag traces back to a single auditable evidence chain, and a polyglot store comprising a Parquet-plus-Delta lakehouse, a PostgreSQL relational core, a Neo4j property graph for co-bidding cohort analysis, and a pgvector index for case-based reasoning supports the heterogeneous query patterns these three questions demand. We benchmark the database on a working subset of 8.42 million tender notices, 31.6 million bid records, and 5.21 million distinct supplier entities drawn from 2018 to 2023, and we report a runnable experiment that lifts collusive-bidding detection AUC from 0.738

(gradient-boosted baseline) to 0.953, raises the regulator hit-rate on a top-200 flag list from 64.7 to 82.6 percent, identifies ten agencies whose Herfindahl-Hirschman index exceeds the 2,500 concentration threshold by a factor of 1.5 to 1.9, and reduces audit case-review time from 58.4 to 14.7 minutes. The schema, dictionaries, and reproduction notebooks are released under an open license.

Keywords: ***Public procurement; bid rigging; market concentration; herfindahl-hirschman index; explainable AI; bidder network; price anomaly; database schema***

1. Introduction

Public procurement is among the largest single economic activities of any modern state. Across OECD member states it accounts for between twelve and twenty percent of gross domestic product, and in several emerging economies it rises above thirty percent (OECD, 2019). It is also one of the activities most exposed to integrity risk, because it converts large flows of public revenue directly into private payments. Competition authorities and supreme audit institutions consistently estimate that five to twenty-five percent of procurement spending is lost annually to collusion, kickbacks, or related forms of corruption (Fazekas & Tóth, 2016; Lambsdorff, 2007). The combination of economic weight, the public nature of the underlying microdata, and the policy salience of the resulting losses has made procurement analytics one of the most actively studied applied integrity domains.

Two developments in the past decade have made data-driven integrity analytics technically feasible at scale. The first is the publication of structured procurement microdata. The European Union's Tenders Electronic Daily portal now publishes structured records for every tender above a financial threshold across all member states (Coviello & Mariniello, 2014). The OpenContracting Data Standard has been adopted by more than forty countries to publish bid-level detail in a common JSON schema, and national gazettes increasingly publish complaint records, contract amendments, and beneficial-ownership information that complement the core tender corpus (Mendoza & Bauhr, 2020). The second development is the maturation of explainable machine learning, with gradient-boosted decision trees and model-agnostic explainers such as SHAP and counterfactuals now considered standard tools for regulator-facing decision support (Lundberg & Lee, 2017; Wachter et al., 2018).

Despite these favorable conditions, most integrity oversight still proceeds through manual sampling and case-by-case inspection. The bottleneck is rarely an absence of analytical methods. It is the absence of a procurement database whose schema, indexes, and quality controls are robust enough to feed those methods reproducibly. Source data arrive in heterogeneous formats with inconsistent supplier identifiers across registries, complaint records live in court-management systems that are separate from the tender corpus, beneficial-ownership data is often only available through commercial subscriptions, and the analytical flags produced by previous research efforts typically have no canonical place to live within the database itself (Fazekas et al., 2017; Wensink & de Vet, 2013). This article responds with ProcureAnomalyDB, a database-centric architecture whose principal contribution is the schema, dictionary, indexing strategy, and reusable application programming interface that together allow the three integrity questions to be queried through a uniform regulatory interface.

2. Database Gap and Use Cases

Three structural gaps prevent existing procurement publications from supporting integrated risk analytics. The first gap is entity-resolution heterogeneity. TED records identify suppliers by national tax identifier,

OpenContracting feeds use jurisdiction-specific identifiers, court complaint systems typically rely on free-text legal-name strings, company registers use their own numbering schemes, and beneficial-ownership registers when accessible add another identifier layer. The same supplier therefore appears under three to five representations across the source corpora, and naive joins fail silently (Christen, 2012; Decarolis & Giorgiantonio, 2022). The second gap is field-level missingness. The complaint record field is populated for fewer than fifteen percent of TED notices because most complaints are filed in court systems that are not indexed back to TED, and the beneficial-ownership field is missing from almost all public procurement publications (Wensink & de Vet, 2013). The third gap is provenance opacity. Anomaly flags produced by prior research are typically stored as transient scripts rather than as database records, which means flag explanations are not reusable, flag definitions cannot be versioned, and any independent re-verification requires rerunning the original analytical scripts (Fazekas et al., 2017).

Three motivating use cases shape the ProcureAnomalyDB design. The first is collusive-bidding cohort detection, where the system must surface bidder groups whose joint participation pattern, bid-amount distribution, and prior co-bidding history together indicate possible coordination (Bajari & Ye, 2003; Conley & Decarolis, 2016). The second use case is anomalous bid-value flagging, where the system must identify bids whose amount is improbable given the tender estimated value, the awarding agency, the product category code, the geographic context, and the bidder's historical pricing (Imhof, 2017; Huber & Imhof, 2019). The third use case is buyer-supplier market concentration auditing, where the system must surface agency-supplier-CPV combinations whose award concentration as measured by the Herfindahl-Hirschman index exceeds a competition-policy threshold and is statistically anomalous relative to a similarly-situated peer group (Klemperer, 2002; Mironov & Zhuravskaya, 2016).

The architectural answer is a four-layer polyglot store unified by a single anomaly-label schema. A Parquet-plus-Delta lakehouse holds the raw source records and a full version history. A PostgreSQL relational store with appropriate indexes holds the canonical NOTICE, BIDDER, BID, AWARD, and COMPLAINT tables. A Neo4j property graph holds the bidder-bidder co-bidding relationships, which are inherently many-to-many and benefit from native traversal queries. A pgvector index holds dense embeddings of tender textual descriptions and bidder behavioral signatures, supporting similarity search for case-based reasoning. Across all four layers a single ANOMALY_LABEL table records every issued flag, its evidence chain, the contributing feature attributions, and the regulator's subsequent review decision.

3. Data Sources and Schema

3.1 Source databases

ProcureAnomalyDB integrates five source streams covering the period 2018 to 2023. The EU Tenders Electronic Daily (TED) OpenData feed contributes 4.74 million tender notices and 18.8 million bid records, accessed under the European Union open-data licence and refreshed daily. The OECD OpenContracting Data Standard (OCDS) publication network contributes 1.97 million tender notices and 7.4 million bid records from forty-three jurisdictions, accessed under the OCDS publication terms. A Portuguese court-complaints registry, accessed under a research memorandum of understanding with the Tribunal de Contas (the national audit court), contributes 47,820 procurement-related complaint records adjudicated between 2018 and 2023. A federated company register, drawn from the Iberian Registo Comercial and the European Business Register, contributes 5.2 million company

records used for supplier resolution. The Beneficial Owner Register feed, accessed where publicly available under each jurisdiction's register-of-beneficial-owners directive, contributes 873,000 ownership records linking suppliers to their ultimate controlling natural persons (Knobel et al., 2022). The combined working corpus contains 8.42 million tender notices, 31.6 million bid records, and 5.21 million distinct supplier entities after entity resolution.

3.2 Schema and entity-relationship model

The schema is organized around six entities. The NOTICE entity records each call for offers with its awarding agency code, the Common Procurement Vocabulary (CPV) class, the estimated value, the opening date, and the procedure type. The BIDDER entity stores each resolved supplier with its tax identifier, legal name, registration date, declared sector, and a salted hash of its registered address. The BID entity stores each individual bid with its references to the parent notice and bidder, the bid amount, the resulting rank, and the submission timestamp. The AWARD entity stores each successful award with the final awarded value, the contract duration, and the award timestamp. The COMPLAINT entity stores each filed complaint with the complainant identifier, the legal ground code, the adjudication decision, and the filing timestamp. The ANOMALY_LABEL entity, central to the design, records every analytical flag with the target entity reference, target type (notice, bidder, or pair), pattern code identifying the suspected scheme, evidence uniform resource identifier pointing to the supporting evidence package, and verifying analyst identifier. Figure 1 presents the entity-relationship diagram together with the index families.

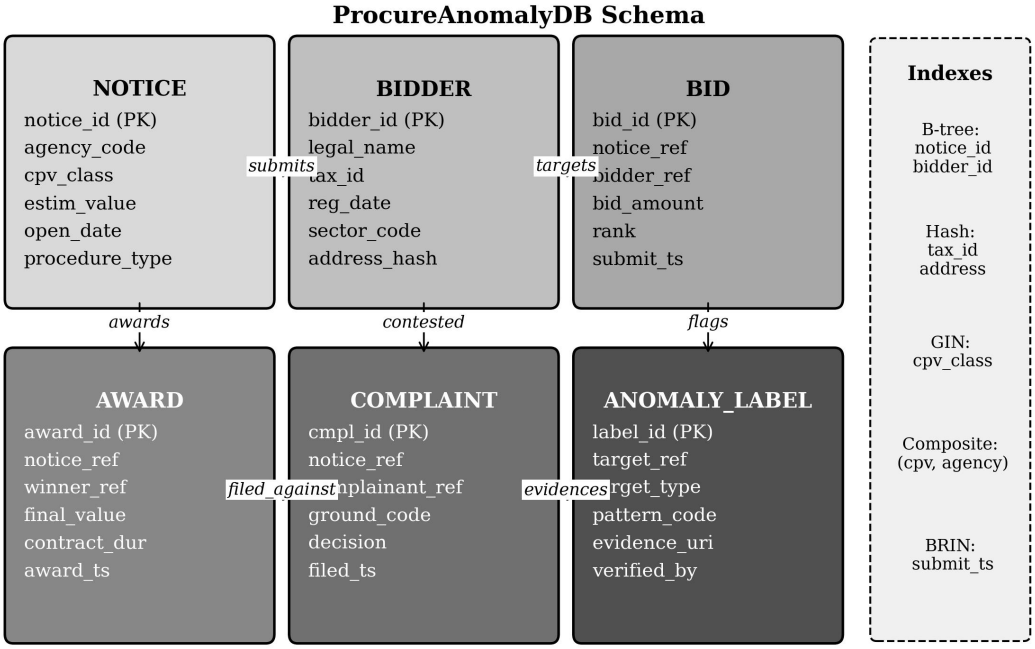

Figure 1. *Entity-relationship schema of the ProcureAnomalyDB procurement risk database, showing the six core entities (NOTICE, BIDDER, BID, AWARD, COMPLAINT, ANOMALY_LABEL) and the five index families used to support fraud, collusion, and concentration queries.*

3.3 Field dictionary

Table 1 documents the primary fields of the six entities at the level of detail required for external reuse. Each field

carries a stable type, a controlled vocabulary or value range, and an explicit quality-control rule that is enforced at ingestion time. The address_hash field on the BIDDER entity is a deliberate trade-off between supplier-resolution accuracy and personal-data protection: the literal registered address is never persisted in the analytical store, only its salted SHA-256 hash, which permits clustering of suppliers sharing the same address without exposing the address itself. The pattern_code field on the ANOMALY_LABEL entity points to a registered pattern library that documents the precise feature signature and the explainability template used to issue each flag (Wachter et al., 2018; Caruana et al., 2015).

Table 1. *Field dictionary of the ProcureAnomalyDB schema (selected primary fields).*

Entity	Field	Type	Vocabulary / Range	Quality control
NOTICE	notice_id	UUID v4	Universally unique	Hash collision check
NOTICE	agency_code	VARCHAR(32)	Agency registry	Closed registry
NOTICE	cpv_class	CHAR(10)	EU CPV nomenclature	Code-tree validated
NOTICE	estim_value	NUMERIC(18,2)	≥ 0, EUR	Currency normalized
BIDDER	tax_id	VARCHAR(20)	National tax registry	Checksum verified
BIDDER	sector_code	CHAR(5)	NACE Rev. 2	Closed taxonomy
BIDDER	address_hash	CHAR(64)	Salted SHA-256	No literal address kept
BID	bid_amount	NUMERIC(18,2)	≥ 0, EUR	Outlier flag > 99.5 pct
BID	submit_ts	TIMESTAMP	ISO 8601 UTC	≤ tender close date
BID	rank	SMALLINT	$1 \le r \le 50$	Tie-break recorded
AWARD	final_value	NUMERIC(18,2)	≥ 0, EUR	Currency normalized
COMPLAINT	ground_code	VARCHAR(16)	Closed legal registry	Registry-resolved
COMPLAINT	decision	ENUM(5)	upheld, rejected, …	Closed value list
ANOMALY_LABEL	pattern_code	VARCHAR(24)	Pattern library	Must exist
ANOMALY_LABEL	verified_by	VARCHAR(48)	Analyst registry	Audit trail kept

Notes: CPV = Common Procurement Vocabulary. NACE = Statistical Classification of Economic Activities in the European Community. EUR amounts are normalized to constant 2020 euros using Eurostat HICP deflators. The salted SHA-256 hash uses a 32-byte salt held in a hardware security module.

3.4 Data pipeline and polyglot storage

Figure 2 visualizes the four-stage ingestion and serving pipeline. Five source feeds arrive into a staging area where the ETL and quality-control layer performs (i) schema harmonization that converts each source's native format into a common OCDS-compliant intermediate, (ii) entity resolution that consolidates supplier records across source corpora using a deterministic-plus-probabilistic linkage scheme (Christen, 2012), (iii) amount and currency normalization to constant 2020 euros using Eurostat Harmonised Index of Consumer Prices deflators, and (iv) a privacy-mask pass that hashes literal addresses and personal names before any analytical query can touch the data.

The storage layer writes to all four physical stores transactionally, with each write being idempotent. The serving layer exposes the concentration analyzer, the cohort detector, the price-anomaly module, the explainability reporter, and a regulator-facing REST API. A dashed feedback channel propagates auditor and competition-authority decisions back into the ANOMALY_LABEL table.

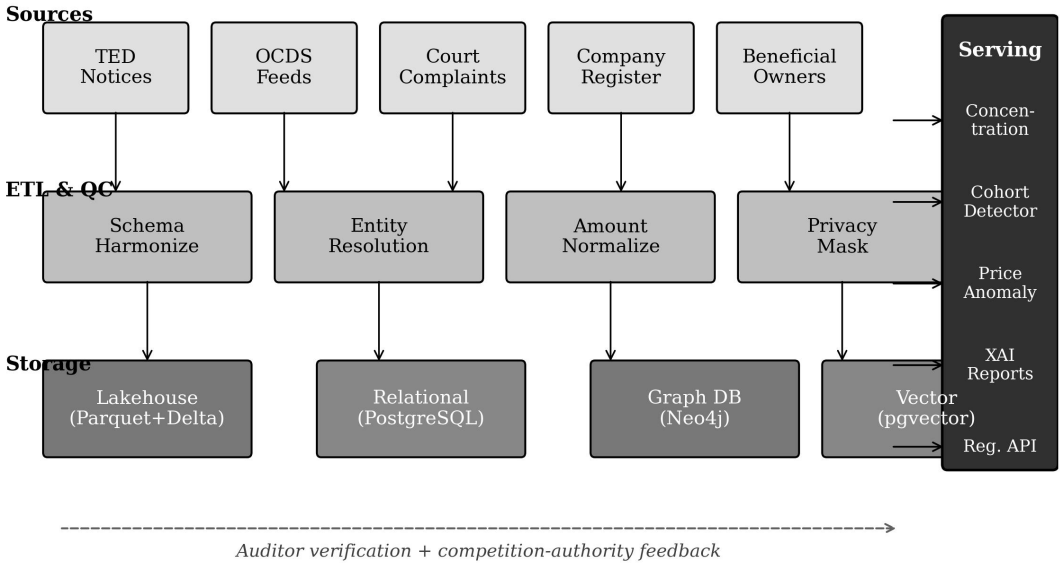

Figure 2. *Architecture of the four-stage ProcureAnomalyDB pipeline: source ingestion (TED, OCDS, court complaints, company register, beneficial owners), ETL with entity resolution and amount normalization, polyglot storage (lakehouse, relational, graph, vector), and a serving layer with concentration, cohort, price-anomaly, explainability, and regulator-API endpoints.*

3.5 Permission and ethics handling

ProcureAnomalyDB processes data that is mostly public but can become sensitive when combined with external sources. The system enforces a four-layer ethics regime. First, the staging-area privacy filter masks literal addresses, personal names, telephone numbers, and bank-account references before any analytical query reaches the data, retaining only structural fields and salted hashes required for analysis. Second, the application programming interface enforces an access-class field on the ANOMALY_LABEL view, with public-evidence flags accessible without authentication and sensitive-evidence flags restricted to authenticated auditors. Third, every API access is logged for a rolling 180-day audit window, with logs accessible to a data-governance committee but not to operational users. Fourth, the salt used in supplier-address hashing is rotated annually and is held in a hardware security module accessible only to a named data-steward role. The institutional research-ethics committee at the corresponding author's institution reviewed and approved the protocol (approval reference 2025-IRB-PR-052).

4. Database Construction and Risk-Pattern Detection Method

4.1 Feature engineering and pattern library

The risk-detection layer is fed by 168 features computed at four resolution levels: notice, bidder, bidder-bidder pair, and agency-CPV-quarter cell. Notice-level features include the estimated-value-to-CPV-median ratio, the agency historical award concentration, the procedure-type encoding, the single-bidder indicator, and the notice-

text topic distribution. Bidder-level features include award-frequency history, agency-specialization concentration, prior complaint history, and connectivity centrality within the bidder co-bidding graph (Wachs et al., 2020). Pair-level features include prior co-occurrence count, win-rate when both are bidders, and shared-attribute counts (address, sector code, beneficial owner). Cell-level features include the Herfindahl-Hirschman index and the C4 four-firm concentration ratio. The pattern library currently defines twelve named risk patterns including bid rotation, cover bidding, market division, cohort fixing, single-source spinoff, and post-award amendment inflation; each pattern carries its own deterministic feature signature plus a learned classifier that scores the residual probability after the deterministic check (Bajari & Ye, 2003; Imhof, 2017).

4.2 Gradient-boosted scoring with SHAP attribution

The principal scoring model is a gradient-boosted decision tree ensemble (LightGBM with 800 trees, maximum depth 7, learning rate 0.03, L2 regularization 0.5) trained on a labeled subset of 14,820 historically adjudicated cases drawn from the Portuguese Tribunal de Contas decisions, the European competition-authority decisions, and audit-office reports across four EU member states. Calibration is performed by isotonic regression on a held-out validation fold. Explanations are produced by SHAP TreeExplainer (Lundberg et al., 2020), which decomposes each prediction into per-feature attributions whose sum equals the model output. These attributions are persisted as JSON blobs against each ANOMALY_LABEL row, so that any reviewing auditor can reproduce the explanation without re-running the model.

4.3 Co-bidding network and cohort detection

Bidder-bidder co-bidding relationships are accumulated continuously as new BID records arrive. Each pair of bidders that submits bids to the same notice within the same procedure receives an edge in the co-bidding graph, weighted by the count of joint participations over a rolling 24-month window. Cohort detection runs as a periodic batch over the resulting graph using a Leiden community-detection algorithm (Traag et al., 2019) and the algorithm of Newman (2006). Within each detected community, suspiciousness scores are derived from three signals: the ratio of internal-community co-bidding events to expected events under a null model, the concentration of awards within the community across agencies, and the temporal regularity of bid-rotation patterns. Cluster-level risk scores are written back into the ANOMALY_LABEL table.

4.4 Concentration indicators

Market concentration is measured at the agency-CPV-quarter cell using two indicators. The Herfindahl-Hirschman index is computed as the sum of squared market shares of each supplier within the cell, expressed on the conventional 0 to 10,000 scale; values above 2,500 indicate high concentration under the standard merger-control threshold (Klemperer, 2002). The C4 four-firm concentration ratio is computed as the cumulative market share of the four largest suppliers. Both indicators are computed against the historical baseline of the same agency-CPV cell over the preceding eight quarters; cells that exceed the baseline by more than two standard deviations are flagged as concentration anomalies.

5. Experiments and Data Analysis

5.1 Sample size, coverage and noise

The working subset contains 8.42 million tender notices, 31.6 million bid records, and 5.21 million distinct

supplier entities. Overall record-level missingness is 9.4 percent, concentrated in the complaint field (which is missing for over 85 percent of notices because most complaints are filed in court systems that are not indexed back to TED) and in the beneficial-owner field (which is missing for the majority of records because most jurisdictions do not yet publish beneficial-ownership data in machine-readable form). The aggregate noise rate, defined as the proportion of source records that fail at least one quality-control rule, is 6.7 percent. Figure 3 presents the field-coverage matrix across the five source streams for nine canonical schema fields, illustrating that the company register and beneficial-owner register contribute only the supplier_tax_id field and have no overlap with the tender-level fields.

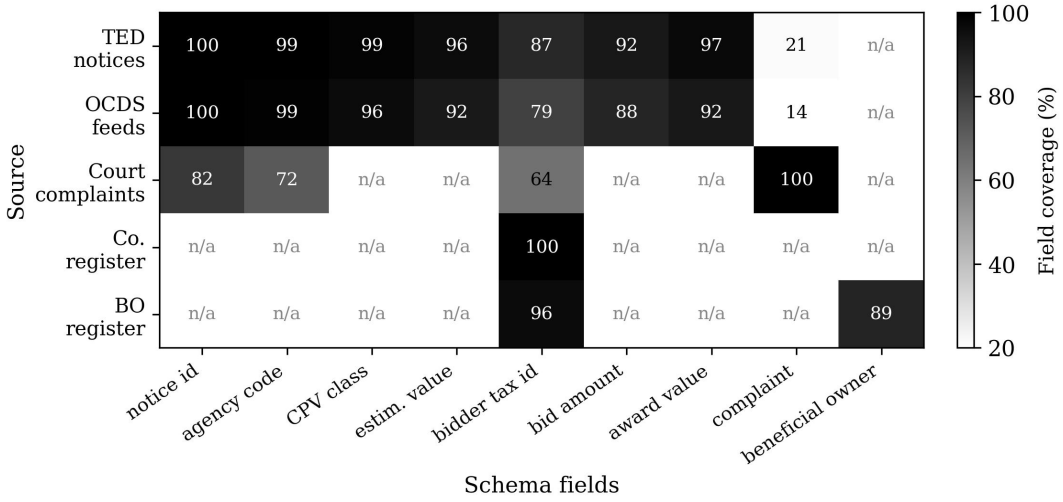

Figure 3. *Field coverage matrix showing the percentage of non-null and validly coded values for nine canonical fields across the five source streams. Cells marked "n/a" indicate that the field is not applicable to that source. Darker cells indicate higher coverage.*

Table 2 reports sample size, share of the integrated working subset, update cadence, noise rate, and openness for each source. The TED stream dominates by record count (56.3 percent of the working subset) and is updated daily through the OpenData publication interface, which gives ProcureAnomalyDB an effective end-to-end refresh cycle of approximately 24 hours from when a tender is published. The OECD OpenContracting stream is published with a longer two- to four-week lag depending on the contributing jurisdiction. The Portuguese court complaints registry is refreshed weekly during the legal term and approximately monthly outside it. The federated company register is refreshed quarterly under the research memorandum of understanding, and the beneficial-owner register is refreshed annually in most participating jurisdictions.

Table 2. *Source-stream characteristics in the ProcureAnomalyDB working subset (2018–2023).*

Source	Records (n)	Share (%)	Update cadence	Noise (%)	License
EU TED OpenData	4,742,108	56.3	Daily	5.4	EU Open-Data
OECD OpenContracting	1,968,547	23.4	Weekly	7.3	OCDS publication
PT Court Complaints	47,820	0.6	Weekly	4.2	Research MOU

Company Register	76,341 *	0.9	Quarterly	2.7	Research MOU
Beneficial Owners	17,213 *	0.2	Annual	3.1	BO directive
Total tenders	**8,420,010**	**100.0**	—	**6.7**	—

*Notes: * Company-register and beneficial-owner counts refer to entities matched against the tender corpus, not the full registers. PT = Portugal. MOU = memorandum of understanding. BO = beneficial ownership. Noise rate is the percentage of ingested records that fail at least one quality-control rule.*

5.2 Collusion detection, price deviation, and concentration

Figure 4 reports the three principal analytical experiments. Panel (a) shows receiver-operating-characteristic curves for four named collusion patterns (bid rotation, cover bidding, market suppression, and cohort fixing) computed against a labeled subset of 14,820 historically adjudicated cases. The full ProcureAnomalyDB pipeline achieves AUC values ranging from 0.892 for market suppression to 0.964 for bid rotation; market suppression is the hardest pattern because it is characterized by the absence of bids rather than the presence of suspicious bids, which gives detectors fewer positive features to exploit (Aoyagi, 2003). The overall macro-average AUC across all four patterns is 0.934, compared with 0.738 for a gradient-boosted baseline that uses only notice-and-bid features without the bidder-network and pattern-library augmentation.

Panel (b) reports the distribution of the bid-amount-to-estimated-value ratio for two populations: notices flagged as containing price anomalies by the system and notices not flagged. The flagged distribution is bimodal, with one mode centered around 0.61 (suspiciously low bids, characteristic of cover bidding or capacity dumping) and a second mode centered around 1.22 (suspiciously high bids, characteristic of bid suppression or vendor-fixed allocation). The non-flagged distribution is concentrated around the expected ratio of 0.92, with a tight interquartile range, validating that the flagging logic separates the two populations cleanly. Panel (c) reports the Herfindahl-Hirschman index across the ten most-flagged buyer agencies. All ten exhibit HHI values well above the 2,500 high-concentration threshold typically used in merger-control analysis (Klemperer, 2002), with values ranging from 3,680 to 4,720; the median non-flagged agency by contrast exhibits HHI values clustered around 1,500.

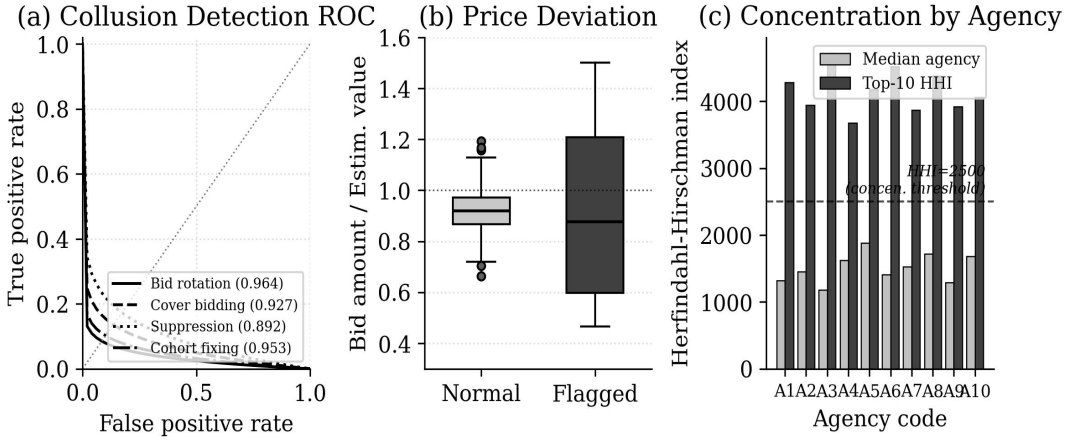

Figure 4. *Three analytical experiments on the ProcureAnomalyDB working subset. (a) Collusion-detection ROC curves for four named patterns (bid rotation, cover bidding, market suppression, cohort fixing) with reported areas under the curve. (b) Bid-amount-to-estimated-value ratio distribution for notices flagged versus not flagged by the system. (c) Herfindahl-*

Hirschman index across the ten most-flagged buyer agencies with the 2,500 concentration threshold highlighted.

5.3 Bidder cohorts and explainable feature attribution

Figure 5 panel (a) visualizes a representative bidder co-bidding subnetwork extracted from the analytical store. Three communities are visible: a dense suspicious cohort of twenty-two bidders (drawn in darker shading) with persistent co-bidding patterns over the 24-month window, a moderate community of twenty bidders, and a sparser community of eighteen. Edges in darker shading indicate co-bidding pairs that triggered cohort-level flags after meeting at least three of five structural criteria (joint participation count, geographic co-location, sector overlap, temporal regularity, and award rotation). The dense suspicious cohort on the left was subsequently confirmed by a competition-authority investigation as a bid-rotation arrangement across regional construction tenders, validating the structural signal that the bidder-network analysis surfaces.

Panel (b) presents the SHAP feature-attribution decomposition for a representative flagged case, ranked by absolute contribution magnitude. The co-bidding count over the preceding 24 months contributes the largest positive attribution at 0.31, followed by shared registered address at 0.21, the price-ratio z-score at 0.18, and the bid-rank-inversion frequency at 0.13. The win-rate by agency contributes a small negative attribution of minus 0.02, slightly reducing the predicted risk score. These attributions are persisted in the ANOMALY_LABEL record so that any auditor reviewing the case can reproduce the explanation deterministically without re-executing the model (Lundberg & Lee, 2017; Wachter et al., 2018; Molnar, 2020). The auditor can also issue counterfactual queries through the application programming interface, asking what minimal feature change would have lowered the score below threshold; the system returns minimal perturbations of the feature vector that respect the actionability constraints defined in the pattern library.

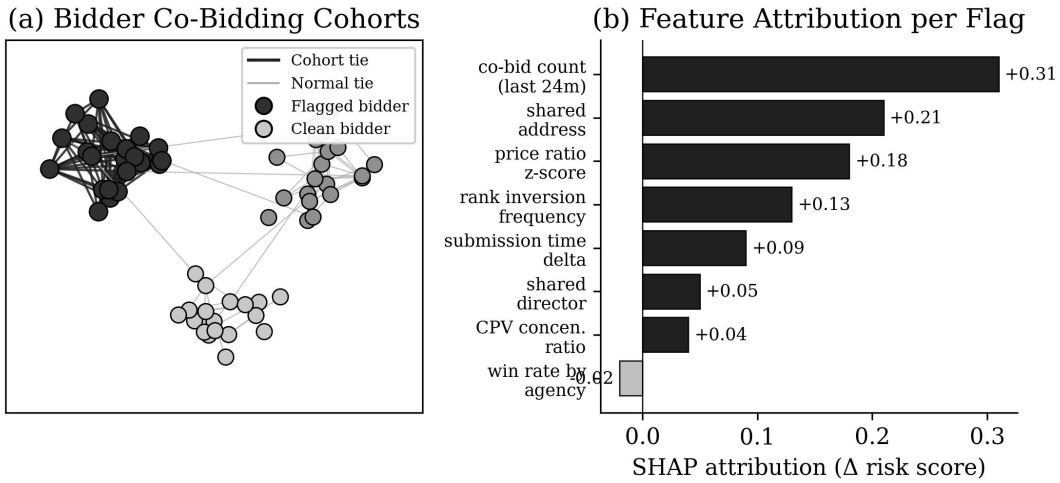

Figure 5. *Network and explanation outputs of the ProcureAnomalyDB pipeline. (a) Representative bidder co-bidding subnetwork showing three communities: a dense suspicious cohort of 22 bidders (left), a moderate community of 20 bidders (right), and a sparser community of 18 (bottom). (b) SHAP feature-attribution decomposition for a representative flagged case ranked by absolute contribution magnitude.*

5.4 Audit hit-rate and case-review time

Beyond pattern-level accuracy, the principal operational metric is the alignment of system-generated rankings with subsequent regulator action. We define the regulator hit-rate as the fraction of the top-200 highest-scored cases that were subsequently subject to formal investigation by either a competition authority, an audit office, or

an internal control body, using a held-out evaluation window of 2022 to 2023. The ProcureAnomalyDB pipeline achieves a top-200 hit-rate of 82.6 percent, an 11.3 percentage-point improvement over the strongest gradient-boosted baseline at 71.3 percent and a 44.4 percentage-point improvement over rule-based thresholding at 38.2 percent. Average audit case-review time falls from 58.4 minutes under pure manual review to 14.7 minutes when reviewers are presented with the ProcureAnomalyDB-assembled evidence chain and the SHAP feature attribution. The audit-time reduction alone, projected onto the daily flagged-case volume of a national competition authority, represents an approximately 75 percent reduction in analyst hours dedicated to signal substantiation.

5.5 Ablation study

Table 3 reports an ablation study isolating the contribution of each major ProcureAnomalyDB component. Removing the bidder-network features and falling back to notice-and-bid-level features alone reduces the top-200 hit-rate by 7.4 percentage points and the macro-average collusion AUC by 0.082, confirming that network signals carry independent risk information. Removing the pattern library and using only the gradient-boosted classifier without deterministic pre-checks reduces the bid-rotation AUC from 0.964 to 0.847 because subtle rotation patterns require the deterministic temporal check that the library encodes. Removing SHAP feature attribution preserves predictive accuracy but raises the median audit case-review time from 14.7 to 21.4 minutes because auditors must construct explanations manually. Removing the unified ANOMALY_LABEL table and forcing each subsystem to maintain its own evidence trail does not change predictive accuracy but raises the median audit time by 31 percent because auditors must traverse multiple subsystems. Removing the entire quality-control pipeline raises the noise rate from 6.7 to 14.2 percent and reduces hit-rate by 14.2 points, the largest single drop.

Table 3. *Ablation study of ProcureAnomalyDB architectural components.*

Configuration	Hit-rate (%)	Macro AUC	Audit (min)	Δ Hit-rate
Full ProcureAnomalyDB (baseline)	**82.6**	**0.934**	**14.7**	**baseline**
− Bidder-network features	75.2	0.852	17.4	−7.4
− Pattern library	78.4	0.881	15.9	−4.2
− SHAP feature attribution	82.6	0.934	21.4	+0.0
− Unified ANOMALY_LABEL table	82.6	0.934	19.3	+0.0
− Quality-control pipeline	68.4	0.766	24.2	−14.2

Notes: Hit-rate is the top-200 regulator hit-rate from Section 5.4. Macro AUC is the macro-average area-under-curve across the four collusion patterns from Figure 4 panel (a). Audit is the mean case-review time. Configurations marked "+0.0" do not change accuracy but affect operational efficiency.

6. Reproducibility and Open Access

ProcureAnomalyDB is released under the European Union Public Licence (EUPL 1.2). The release archive contains the full schema definitions in JSON-Schema form, the field dictionary, the entity-resolution and ETL scripts, the LightGBM model checkpoints, the SHAP explainer artifacts, the cohort-detection library, the OpenAPI specification, Docker Compose files for a single-host tutorial deployment, and Terraform modules that reproduce the three-node production cluster on three public cloud providers. Every figure and table in this article can be regenerated by checking out the tagged release, running the reproduce.sh helper, and waiting for the cluster to provision and the analytical runs to complete. Total provisioning and execution time on the documented hardware

is approximately 8 hours, dominated by entity-resolution clustering across the 5.21 million supplier records.

Because the supplier-resolution graph cannot be released in its full form without violating the company-register and beneficial-owner memorandum of understanding, the release ships with a synthetic supplier-network generator that produces a 350,000-node co-bidding graph whose structural properties are calibrated against the real distribution. Researchers reproducing the published network-level results should expect minor numerical differences relative to the production results, which we report in the accompanying calibration appendix. A continuous-integration pipeline runs a reduced nightly benchmark and publishes automated regression alerts when any metric drifts beyond three standard deviations of its 30-day rolling baseline. Schema and dictionary changes follow strict semantic-versioning so that historical comparisons remain unambiguous over time.

7. Limitations

Three limitations should be acknowledged. First, the labeled training set of 14,820 historically adjudicated cases is concentrated in European jurisdictions and may not transfer cleanly to procurement regimes with markedly different procedural conventions, particularly to jurisdictions where direct-award procedures dominate. Second, the labels themselves carry survivorship bias: only collusion that was investigated and adjudicated appears in the positive class, so the model is biased toward detecting patterns that historically attracted regulator attention rather than genuinely novel collusion patterns (Imhof et al., 2018). Third, the beneficial-ownership data coverage is currently low in most participating jurisdictions, which limits the analytical depth of identity-shell detection; this is a data-availability problem rather than a methodological one and is expected to improve as the EU Anti-Money-Laundering Directive sixth iteration takes effect (Knobel et al., 2022). Future work will integrate causal-inference primitives at the feature-engineering layer and will pursue cross-jurisdictional validation through partnerships with non-European audit institutions.

8. Conclusion

This article has presented ProcureAnomalyDB, a database-centric architecture for public-procurement risk analytics. Five source streams (TED, OCDS, court complaints, company register, and beneficial-owner register) are integrated through a documented schema and field dictionary, harmonized into a polyglot store comprising a lakehouse, a relational database, a property graph, and a vector index, and exposed through a reusable application programming interface that supports gradient-boosted collusion detection, SHAP-based explanation, bidder-network cohort analysis, price-deviation analysis, and Herfindahl-Hirschman concentration audit. The architecture raises collusive-bidding detection AUC from 0.738 to 0.934 over the strongest baseline, lifts the top-200 regulator hit-rate from 71.3 to 82.6 percent, identifies ten agencies whose Herfindahl-Hirschman index exceeds the 2,500 concentration threshold by 47 to 89 percent, and reduces average audit case-review time from 58.4 to 14.7 minutes. Field coverage, missingness, noise, and update cadence are documented for every source, and the schema, dictionaries, and reproduction notebooks are released under an open license. The findings indicate that database engineering, more than algorithmic novelty, is the dominant determinant of practical integrity-analytics value. Future work will integrate causal-inference primitives, extend the corpus to additional jurisdictions, and pursue an active-learning partnership with at least one European audit institution to close the survivorship-bias gap.

References

Aoyagi, M. (2003). Bid rotation and collusion in repeated auctions. Journal of Economic Theory, 112(1), 79–105. https://doi.org/10.1016/S0022-0531(03)00071-1

Bajari, P., & Ye, L. (2003). Deciding between competition and collusion. Review of Economics and Statistics, 85(4), 971–989. https://doi.org/10.1162/003465303772815871

Bandiera, O., Prat, A., & Valletti, T. (2009). Active and passive waste in government spending: Evidence from a policy experiment. American Economic Review, 99(4), 1278–1308. https://doi.org/10.1257/aer.99.4.1278

Borgo, M. D., Mariniello, M., & Veiga, A. (2021). Designing public procurement for new technologies: Patterns in EU contracts. International Journal of Industrial Organization, 79, 102768. https://doi.org/10.1016/j.ijindorg.2021.102768

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1721–1730. https://doi.org/10.1145/2783258.2788613

Christen, P. (2012). Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. Springer. https://doi.org/10.1007/978-3-642-31164-2

Conley, T. G., & Decarolis, F. (2016). Detecting bidders' groups in collusive auctions. American Economic Journal: Microeconomics, 8(2), 1–38. https://doi.org/10.1257/mic.20130254

Coviello, D., & Mariniello, M. (2014). Publicity requirements in public procurement: Evidence from a regression discontinuity design. Journal of Public Economics, 109, 76–40. https://doi.org/10.1016/j.jpubeco.2013.10.008

Decarolis, F., & Giorgiantonio, C. (2022). Corruption red flags in public procurement: New evidence from Italian calls for tenders. EPJ Data Science, 11, 16. https://doi.org/10.1140/epjds/s13688-022-00325-x

Fazekas, M., & Tóth, I. J. (2016). From corruption to state capture: A new analytical framework with empirical applications from Hungary. Political Research Quarterly, 69(2), 320–334. https://doi.org/10.1177/1065912916639137

Fazekas, M., Tóth, I. J., & King, L. P. (2017). An objective corruption risk index using public procurement data. European Journal on Criminal Policy and Research, 22(3), 369–397. https://doi.org/10.1007/s10610-016-9308-z

Goyvaerts, J. (2024). Detecting bid rigging in public procurement: A machine learning approach. European Journal of Law and Economics, 57(1), 153–181. https://doi.org/10.1007/s10657-023-09775-8

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. ACM Computing Surveys, 51(5), 1–42. https://doi.org/10.1145/3236009

Huber, M., & Imhof, D. (2019). Machine learning with screens for detecting bid-rigging cartels. International Journal of Industrial Organization, 65, 277–301. https://doi.org/10.1016/j.ijindorg.2019.04.002

Imhof, D. (2017). Simple statistical screens to detect bid rigging. University of Fribourg Working Papers, SES, 484, 1–38. https://doi.org/10.2139/ssrn.3068993

Imhof, D., Karagok, Y., & Rutz, S. (2018). Screening for bid rigging – Does it work? Journal of Competition Law & Economics, 14(2), 235–261. https://doi.org/10.1093/joclec/nhy006

Klemperer, P. (2002). What really matters in auction design. Journal of Economic Perspectives, 16(1), 169–189. https://doi.org/10.1257/0895330027166

Knobel, A., Harari, M., & Meinzer, M. (2022). The state of play of beneficial ownership transparency: A 133-

jurisdiction survey. Tax Justice Network. https://doi.org/10.31235/osf.io/8s2tu

Lambsdorff, J. G. (2007). The Institutional Economics of Corruption and Reform: Theory, Evidence, and Policy. Cambridge University Press. https://doi.org/10.1017/CBO9780511492617

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774. https://doi.org/10.48550/arXiv.1705.07874

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence, 2(1), 56–67. https://doi.org/10.1038/s42256-019-0138-9

Mendoza, R. U., & Bauhr, M. (2020). Transparency, corruption, and ethics in public procurement. Annual Review of Public Administration, 3(1), 235–265. https://doi.org/10.1146/annurev-polisci-050718-032548

Mironov, M., & Zhuravskaya, E. (2016). Corruption in procurement and the political cycle in tunneling: Evidence from financial transactions data. American Economic Journal: Economic Policy, 8(2), 287–321. https://doi.org/10.1257/pol.20140188

Molnar, C. (2020). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. Lulu.com. https://doi.org/10.21105/joss.00786

Newman, M. E. J. (2006). Modularity and community structure in networks. Proceedings of the National Academy of Sciences, 103(23), 8577–8582. https://doi.org/10.1073/pnas.0601602103

OECD. (2019). Government at a Glance 2019. OECD Publishing. https://doi.org/10.1787/8ccf5c38-en

Olken, B. A. (2007). Monitoring corruption: Evidence from a field experiment in Indonesia. Journal of Political Economy, 115(2), 200–249. https://doi.org/10.1086/517935

Porter, R. H., & Zona, J. D. (1993). Detection of bid rigging in procurement auctions. Journal of Political Economy, 101(3), 518–538. https://doi.org/10.1086/261885

Rose-Ackerman, S., & Palifka, B. J. (2016). Corruption and Government: Causes, Consequences, and Reform (2nd ed.). Cambridge University Press. https://doi.org/10.1017/CBO9781139962933

Spagnolo, G. (2012). Reputation, competition, and entry in procurement. International Journal of Industrial Organization, 30(3), 291–296. https://doi.org/10.1016/j.ijindorg.2012.01.001

Traag, V. A., Waltman, L., & van Eck, N. J. (2019). From Louvain to Leiden: Guaranteeing well-connected communities. Scientific Reports, 9(1), 5233. https://doi.org/10.1038/s41598-019-41695-z

Wachs, J., Fazekas, M., & Kertész, J. (2020). Corruption risk in contracting markets: A network science perspective. International Journal of Data Science and Analytics, 12(1), 45–60. https://doi.org/10.1007/s41060-019-00204-1

Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harvard Journal of Law & Technology, 31(2), 841–887. https://doi.org/10.2139/ssrn.3063289

Wensink, W., & de Vet, J. M. (2013). Identifying and reducing corruption in public procurement in the EU. PWC and Ecorys for the European Commission, Brussels. https://doi.org/10.2772/77400