

Detecting Shill Bidding in eBay Auctions: Evidence from Collectible Coins

Seth Kushniryk

Department of Economics, Queen's University

October 19, 2026

Abstract

Shill bidding, the practice of sellers using fake identities to bid on their own auctions, persists on online auction platforms despite explicit prohibitions. This paper asks two questions: can shill bidding be detected from observable bidding behaviour alone, and what effect does it have on final prices? Using bid histories for 2,520 eBay auctions of certified graded collectible coins, of which 1,561 enter the analysis sample, I construct six behavioural metrics from the bid sequences (none uses auction price as an input) and apply hierarchical clustering to flag auctions consistent with shill activity. Because the seller is the one shilling, treatment is not random across the platform: the sellers who shill differ from those who do not, and within a given seller the choice of which auctions to shill is also non-random. I report two complementary estimates. The broader between-seller comparison, on the full sample with grade fixed effects, yields a price premium of 17.4 percentage points of book value. A narrower within-seller comparison, restricted to the 35 sellers in the sample whose portfolios contain both flagged and non-flagged auctions, yields 22.8 percentage points ($p < 0.0001$, $n = 1,231$). That the within-seller estimate is larger suggests observable seller selection was biasing reduced-form estimates downward.

Legal disclaimer: This paper identifies bidding patterns statistically consistent with shill bidding using publicly available eBay auction data. All identifying information (including bidder usernames, seller usernames, and auction listing numbers) has been anonymised. The classification of bidder-auction pairs as “suspicious” is based entirely on unsupervised clustering of behavioural metrics. No allegations of shill bidding have been proven in a court of justice. For the purpose of this analysis, I treat the identified patterns as shill bidding.

1 Introduction

In 2023, eBay facilitated \$73.2 billion in gross merchandise volume across billions of auction and fixed-price listings (eBay Inc., 2024). Feedback scores and seller-reputation mechanisms have curbed most forms of fraud on the platform; shill bidding has proved harder to dislodge. Shill bidding occurs when sellers bid on their own auctions, directly through alternate accounts or via a confederate, inflating prices and manipulating the auction process. The shill bidding literature has documented three distinct strategies: bids placed early to signal demand, bids placed to lift the standing price above an undisclosed reserve, and incremental bids that probe the winner’s hidden proxy ceiling (Chakraborty & Kosmopoulou, 2004; Engelberg & Williams, 2009; Lamy, 2009). A UK Office of Fair Trading survey calls shill bidding a “common problem” on online auction platforms (OFT 2007, cited in Marshall 2024); Kauffman and Wood (2003) note that the practice is treated as illegal in many jurisdictions, including the United Kingdom and the United States. Beyond eBay, where users have been criminally prosecuted for the practice, allegations and suggestive evidence have been documented at the vehicle platform Copart, the real-estate platform Auction.com, and the penny-auction site PennyBiddr, the last of which was shut down following such allegations (Marshall, 2024). eBay itself operates an automated “shill hunter” system but does not publicly disclose detection methodology or confirmed cases, leaving researchers without ground-truth labels (a verified roster of confirmed shill auctions that a supervised classifier could be trained against).

This paper asks two questions. First, can shill bidding be detected from observable bidding behaviour alone? Secondly, given a classifier that does not use price information, what effect does shill bidding have on final prices? The empirical literature on shill bidding has shed some light on both but is sparse, constrained by the joint absence of ground-truth labels and platform cooperation. Engelberg and Williams (2009) infer a 1.39% bid-level shill rate on eBay event tickets. Marshall (2024) develops identification strategies for English auctions using eBay auction data on Armani Acqua di Gio perfume. Both adopt reduced-form bounding designs because, absent ground-truth labels, there is little else the data will support: a structural model of shill bidding would require parametric assumptions about shill strategies that no observational sample can discipline. Parallel work in computer science has applied hierarchical clustering to hand-engineered behavioural features (Alzahrani & Sadaoui, 2018, 2020; Anowar, Sadaoui, & Mouhoub, 2018), producing classifications but stopping short of using them as treatments in causal estimation.

I take a third route. The methodologically new contribution is the use of a behavioural

classifier as the treatment variable in a multi-design causal analysis: I construct six behavioural metrics from the bid sequences (none uses auction price as an input), cluster bidder-auction pairs in the resulting feature space, and use the auction-level classification as a treatment variable in subsequent causal estimation. The metric set itself is largely inherited. It adopts five of Alzahrani and Sadaoui (2018)’s eight bidder-level markers, computed from raw eBay bid histories: Bidder Tendency (a bidder’s concentration on specific sellers), Early Bidding (the timing of the first bid relative to auction start), Successive Outbidding (the count of consecutive proxy-outbid bids by one bidder), Bidding Ratio (the bidder’s share of bids within an auction), and Winning Ratio (the fraction of entered auctions won). I make two refinements: the binary Successive Outbidding indicator is replaced by a sequence-length measure that captures the price-squeezing dynamics described by Engelberg and Williams (2009), and the bidder-auction Early Bidding metric is aggregated to the bidder level. Marshall’s identification rests on auction-mechanism restrictions in English auctions; mine rests on a design-based comparison conditional on a behavioural classification. Section 4.3 traces one flagged auction step by step to illustrate the price-squeezing signature that the clustering picks up.

I focus on a specific eBay category, certified graded collectible coins, chosen because the setting offers three features that prior empirical work has lacked. First, certified coins are catalogued by Krause–Mishler reference numbers (KM#) and graded on standardised PCGS/NGC numeric scales, so any two auctions for the same KM# at the same grade involve essentially the same product. Secondly, NGC publishes book values constructed independently of any individual eBay auction, making sale-to-book a price-independent normalisation that does not contaminate downstream causal inference. Thirdly, the scraped sample covers 462 distinct coin types (by Krause–Mishler reference number) issued by 130 countries within a single market segment, providing within-coin homogeneity and cross-coin heterogeneity adequate for fixed-effects estimation. Using a custom web scraper, I collected detailed bid histories for 2,520 completed auctions, capturing the complete sequence of bids, anonymised bidder identifiers, bid timing, and bidder feedback scores.

With suspicious auctions identified (144 of the 1,561 sample auctions, or 9.2%, contain at least one bidder-auction pair flagged by the clustering procedure), the second question is what effect shilling has on final prices. The main empirical challenge here is that the seller is the principal who orchestrates any shilling, whether bidding through alternate accounts directly or through a confederate, so treatment is not assigned randomly across the platform. Two layers of selection follow. Between-seller selection: the sellers who shill differ from those who do not on many dimensions, so a naive comparison conflates the shill effect with seller heterogeneity. Within-seller selection: of the 172 sellers in the

sample, 35 have at least one flagged auction in a portfolio that also contains non-flagged auctions, and the choice of which auction in such a portfolio to shill is unlikely to be random either; the auctions chosen for shilling may differ in expected demand, item value, or other characteristics from the rest of the seller’s portfolio.

I address the two layers with complementary designs. The broader between-seller comparison uses the full sample with grade fixed effects and propensity-score matching; this delivers the largest sample but cannot separate the shill effect from time-invariant seller heterogeneity. The narrower within-seller comparison restricts attention to the 35 mixed-portfolio sellers and asks how much higher a given seller’s flagged auctions sell than that same seller’s non-flagged auctions. This neutralises between-seller selection but leaves within-seller auction-level selection in place. Exact-match designs at the seller-coin-grade level shrink the room for within-seller selection on observables; Oster (2019) bounds quantify how strong unobservable selection would have to be to nullify the result. Experimental work brackets the magnitude: Nikitkov and Bay (2015) documents 16–44% premia in auctions with experimentally controlled shill bidding, and McCannon and Minuci (2020) and Ingebretsen Carlson and Wu (2022) characterise buyer responses to the threat of shilling in laboratory auctions.

The broader between-seller comparison with grade fixed effects yields 17.4 percentage points of book value ($p = 0.016$, $n = 1,558$). The within-seller specification yields 22.8 ($p < 0.0001$, $n = 1,231$). That absorbing seller heterogeneity *raises* rather than eliminates the estimate is evidence that observable seller selection had been biasing reduced-form estimates downward. Estimates cluster between 17 and 25 percentage points across specifications that absorb seller heterogeneity; the tightest exact-coin-within-seller match (65 pairs) collapses to 3.4 percentage points and cannot reject a null. Oster (2019) bounds on the within-seller specification suggest that nullifying the within-seller estimate would require unobservable auction-level selection many times stronger than, and opposite-signed to, observable selection. The estimates fall in the lower-middle of the 16–44% experimental range.

The paper bridges the computer-science fraud-detection literature on unsupervised behavioural clustering with the economic-identification literature on shill bidding, treating the cluster classification as a treatment variable in design-based causal estimation. From that pairing comes the first within-seller estimate of the shill-bidding price premium, and the finding that observable selection on seller characteristics has been masking the effect downward in prior reduced-form estimates. Because the detection metrics do not use price information, the same hierarchical clustering can in principle run in real time against the live bid stream, a platform-governance lever that prior identification-based

approaches do not afford.

2 Background

2.1 Shill Bidding: Theory, Types, and Effects

Three forms of shill bidding are recognised in the literature, each with a different observable signature.

Reserve-price shilling. The seller submits bids to push the standing price above an undisclosed reserve, effectively implementing a secret reserve through fraudulent bidding rather than the platform’s official mechanism. Motivations include avoiding platform fees levied on stated reserves and avoiding the deterrent effect that visible reserves have on bidder entry (Bajari & Hortaçsu, 2003). The defining feature is that the shill bids cease once the standing price clears the desired floor.

Signal-effect shilling. The seller (or accomplice) submits bids early in the auction to create an appearance of demand or competition, potentially attracting genuine bidders who interpret the activity as a positive signal about the item’s value or quality. This is most relevant in common-value or interdependent-value settings, where legitimate bidders update their beliefs from observed bidding activity (Lamy, 2009). The defining feature is timing: signal-effect shill bids are concentrated in the early portion of the auction, before legitimate bidders have committed.

Price-squeezing shilling. The seller (or accomplice) exploits eBay’s proxy-bidding mechanism by submitting incremental bids to probe the current high bidder’s hidden proxy maximum without winning. Each shill bid prompts an automatic proxy outbid; the shill increments to the next standard increment and prompts another, ratcheting the price upward. The shill stops just before the high bidder’s maximum is exceeded, so the price equals (rather than falls short of) that maximum (Engelberg & Williams, 2009). Barbaro and Bracht (2021) characterise a closely related strategy that exploits bid retraction rules: the seller submits high shill bids that reveal buyers’ proxy maximums through eBay’s automatic bidding system, then retracts them and resubmits a bid matching the revealed valuation. The defining feature is a sequence of small incremental bids by a single bidder ending with a sudden gap or withdrawal. A worked example appears in Table 4, and a flagged auction is traced bid by bid in Section 4.3.

Each strategy leaves a different observable trace, and the detection metrics constructed in Section 3.2 are designed to capture all three: timing-based markers target signal-effect,

sequence-based markers target price-squeezing, and cross-cutting markers capture the seller-bidder coordination common to all variants. A priori, any of the three types could dominate in a given sample, so the metric set is included in full to remain faithful to the literature’s taxonomy. The clustering algorithm in Section 3.3 combines these markers into a single suspicious/normal classification; which combination of marker types actually drives the suspicious cluster in this sample is an empirical question, taken up in the next paragraph and in Section 4.

In the empirical results, however, the suspicious cluster recovered by the clustering algorithm is dominated by price-squeezing markers and the cross-cutting coordination markers, with timing-based (signal-effect) markers showing the smallest separation between flagged and unflagged bidders (Section 4). One reading is that platform-outsider data favour some shill types over others. Reserve-price shilling is defined by bids ceasing once an undisclosed reserve clears, a threshold no outsider can observe, so this variant is essentially undetectable in public bid data. Signal-effect shilling is in principle detectable through the joint behaviour of seller-concentration and early-bidding markers, but in a category like collectible coins, where many honest bidders place early bids to lock in proxy maximums or simply register interest, early bids by enthusiasts are difficult to distinguish from early shill bids on timing alone, so the timing-based markers add less separation than the cross-cutting markers. Price-squeezing, by contrast, leaves a mechanically deterministic trace: a sequence of standard-increment proxy outbids ending in gap compression, dictated by eBay’s proxy mechanism rather than by bidder discretion. The methodology in this paper is therefore most diagnostic for price-squeezing in combination with cross-auction coordination fingerprints; reserve-price shilling is not identifiable with these data, and signal-effect shilling, if present, does not separate from legitimate early bidding in this sample.

Theory predicts a positive but bounded effect of shill bidding on prices. In an honest second-price (or eBay-style proxy) auction, the price equals the second-highest valuation: the winner pays just enough to outbid the runner-up (Vickrey, 1961); Milgrom (2004) provides a textbook treatment. A shill bidder probing the winner’s hidden proxy maximum changes this. Each successive shill bid prompts the winner’s proxy to outbid by one increment; if the shill stops just below the winner’s maximum, the price equals (rather than falls short of) that maximum. The premium therefore equals the gap between the winner’s true valuation and the second-highest legitimate valuation, a quantity bounded above by the dispersion of valuations among legitimate bidders. Under signal-effect motives, additional bids can also raise legitimate bidders’ beliefs about the item’s value, further inflating prices (Lamy, 2009). The exception is the common-value case of Kosmopoulou

and De Silva (2007), in which sophisticated bidders anticipating shilling shade their bids downward to mitigate the winner’s curse, a force that can offset and in principle reverse the price-raising effect.

Experimental evidence bounds the magnitude. Nikitkov and Bay (2015) documents 16–44% price premia in auctions with experimentally controlled shill bidding. McCannon and Minuci (2020) show a “shiller’s curse” in laboratory auctions: buyers reduce bids in anticipation of shill activity and seller revenue falls by 4.7% on average across auction formats (7.9% in second-price auctions). Ingebretsen Carlson and Wu (2022) extend this to sequential auctions and find that the mere threat of shilling raises early-auction bids, even when shilling does not in fact occur.

2.2 Identification and Detection in Observational Data

Identifying shill bidding empirically poses two challenges. Detection is hard because the practice is by design covert: platforms do not label bids as fraudulent, sellers actively mimic legitimate patterns, and standard supervised learning is precluded by the absence of ground-truth labels. Causal estimation is hard because even given a classifier, sellers may shill selectively, so any naive comparison of treated and control auctions confounds the shill effect with selection on item or seller characteristics. A further tension is that prices serve dual roles as the outcome of interest and as a potentially informative signal for detection: using price information as input to detection risks creating mechanical correlation when prices subsequently serve as the dependent variable.

Existing strategies have addressed these challenges in two ways. Standard structural estimation in empirical auction analysis makes strong assumptions about observed bids: Guerre, Perrigne, and Vuong (2000) assumes all observed bids represent genuine demand, whereas Hickman, Hubbard, and Paarsch (2017) and Hickman, Boehnke, and Bodoh-Creed (2021) require this only of the two highest bids and abstract from modelling the specific strategies that generated the other observed bids. No structural model that explicitly accommodates shill bidding has been estimated empirically, because the parametric demands on shill strategies and bidder beliefs exceed what observational data can discipline. Marshall (2024) provides partial-identification bounds for English auctions, deriving conditions under which the distribution of valuations is identifiable even when the seller engages in shill bidding. Field experiments with controlled shill bidding (Nikitkov & Bay, 2015) offer strong causal identification but face ethical and scale limitations.

For detection without ground truth, the computer-science fraud-detection literature has applied unsupervised hierarchical clustering on hand-engineered behavioural fea-

tures. Alzahrani and Sadaoui (2018) introduce a set of bidder-level markers (Bidder Tendency, Successive Outbidding, Bidding Ratio, Winning Ratio, Early Bidding) computed from eBay bid histories. Alzahrani and Sadaoui (2020) review clustering-based methods for auction fraud detection, emphasising hierarchical algorithms; the Clustering Using REpresentatives (CURE) algorithm (Guha, Rastogi, & Shim, 1998) represents clusters by several well-scattered points rather than a single centroid and is the workhorse adopted here. Anowar et al. (2018) extend the clustering framework to multi-stage detection. The output of these methods is a classification, not a causal estimate; pairing them with a design-based estimator is what this paper contributes.

3 Data and Methodology

3.1 Data Collection

Product Selection and Sample Frame

This study focuses on eBay auctions for certified graded collectible coins. The category has three properties that suit the empirical exercise: certified grading by professional services delivers near-standardised product characteristics and a numerical grade that doubles as a fixed effect for quality; liquid secondary markets and published price guides supply value benchmarks independent of any eBay auction outcome; and the average winning bid in my sample exceeds \$100, large enough to make manipulation worth the seller’s trouble.

The sample consists of 2,520 completed auctions for coins with professional grades and Krause–Mishler catalogue numbers (KM#). Auctions ran for 1, 3, 5, 7, or 10 days; the duration variation matters because it enters the denominator of the timing-based metrics.

Web Scraping Implementation

I developed custom Python scripts using Selenium for automated browsing and BeautifulSoup 4 for HTML parsing to extract detailed auction data from eBay. For each completed auction, I collected the full bid history: the sequence and timing of all bids, anonymised bidder identifiers assigned by eBay, bid amounts and current auction prices after each bid, bidder feedback scores at the time of bidding, auction characteristics (starting price, duration, final sale price), and seller identifiers.

eBay preserves the complete bid record for some period after an auction closes, but the pages are dynamically loaded and paginated; Selenium handles the navigation and Beau-

tiful Soup extracts the structured fields from the resulting HTML. Preprocessing removes duplicate records, harmonises timestamps, and assigns unique auction identifiers.

Control Variables and Book Values

For a measure of item value independent of auction outcomes, I take book values from the Numismatic Guaranty Company (NGC) Price Guide, the standard reference among certified-coin collectors. NGC publishes expert price estimates at each grade level. The NGC guide is a modern, more comprehensive equivalent to the *Coins Magazine* surveys used by Bajari and Hortaçsu (2003) in their study of internet auctions for collectible coins. These serve both as the normalising denominator for auction prices (enabling comparison across items of different intrinsic worth) and as a control in the regressions.

For some coin-grade combinations, NGC price guide values were unavailable. In these cases, I imputed book values using linear interpolation between adjacent available grades or, when necessary, used the nearest available grade's value. Approximately 15% of auctions required some form of book value imputation.

Sample and Summary Statistics

After restricting to auctions with certified grades, excluding private listings, and retaining only successful sales with complete bid histories, the analysis sample comprises 1,561 of the 2,520 collected auctions. Focusing on completed auctions introduces a potential survivorship concern: if shill bidding systematically affects the probability of sale, conditioning on successful transactions would bias effect estimates. The detection methodology described below identifies suspicious behaviour from bidding patterns rather than sale outcomes, breaking the most direct mechanical correlation; sample retention rates are also consistent with broader eBay completion rates reported by Bajari and Hortaçsu (2003).

Table 1 presents auction-level summary statistics. Book values and winning bids are both right-skewed, with means roughly double medians; both range over four orders of magnitude. Auctions attract a mean of 22.7 bids from 6.5 unique bidders, implying that the average bidder places 3–4 bids, consistent with eBay's proxy bidding mechanism, which automatically increments offers up to a stated maximum (Engelberg & Williams, 2009).

Figures 1–4 display the distributions of book value, winning bid, bid count, and bidder count. Book values cluster in the \$0–\$100 range, winning bids in the \$25–\$75 range. Bid counts are approximately log-normal with a long right tail to 131; bidder counts are rightward-skewed with a mode around 5–8.

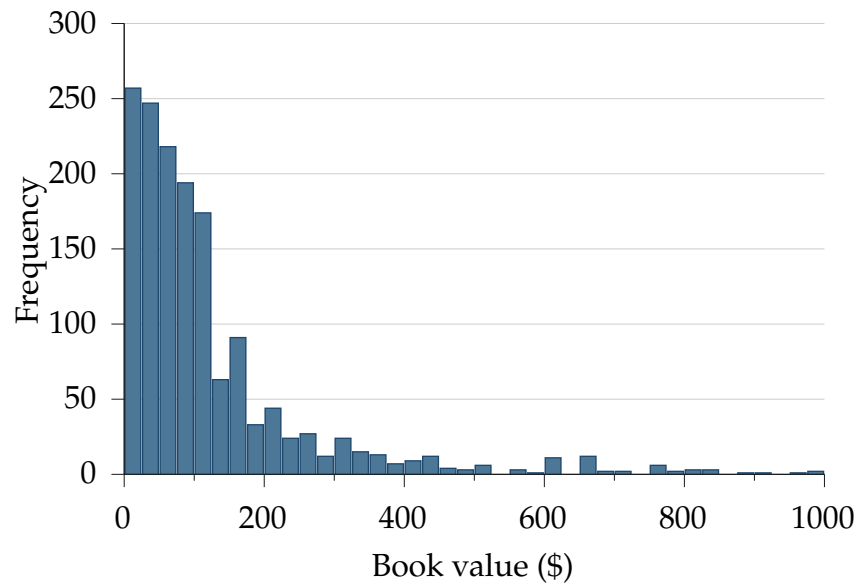


Figure 1: Distribution of Book Values (Under \$1,000; 1,527 of 1,559 auctions)

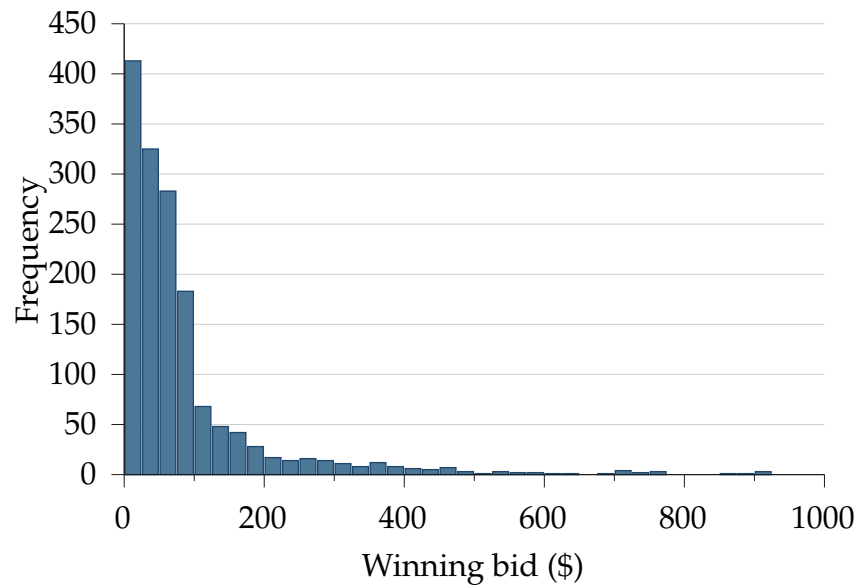


Figure 2: Distribution of Winning Bids (Under \$1,000; 1,536 of 1,561 auctions)

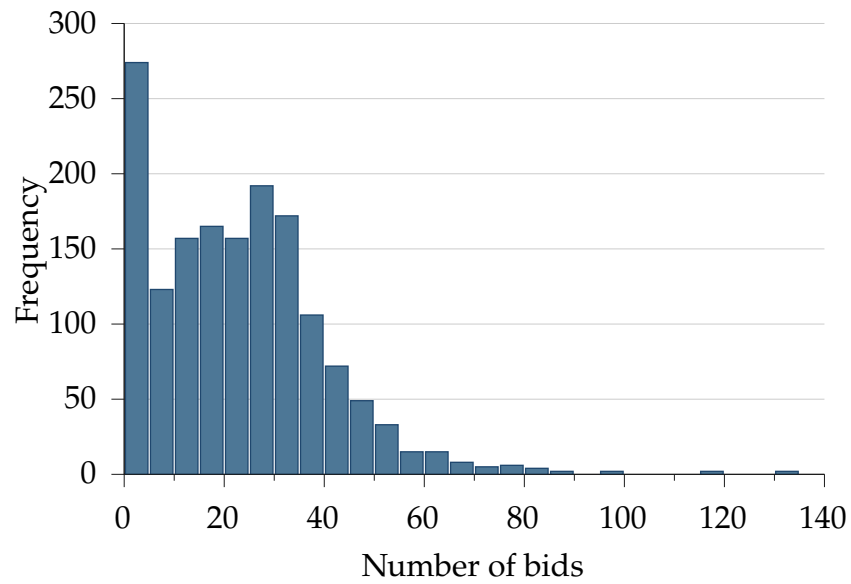


Figure 3: Distribution of Number of Bids per Auction

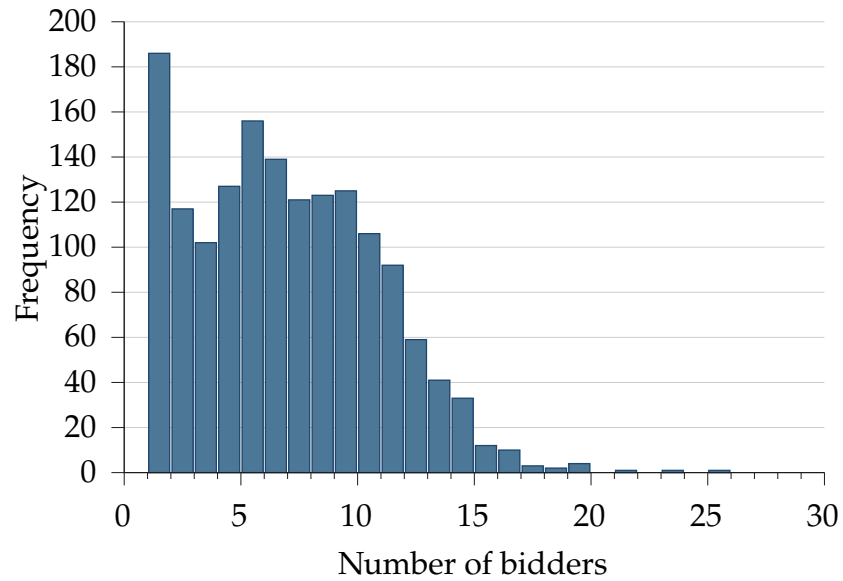


Figure 4: Distribution of Number of Bidders per Auction

Table 1: Summary Statistics for Auction-Level Variables

Variable	N	Mean	Median	Std. Dev.	Min	Max
Book Value (\$)	1,559	164.93	80.00	388.21	0.30	5,079.22
Winning Bid (\$)	1,561	114.15	52.50	276.91	0.45	3,605.55
No. of Bids	1,561	22.72	22	17.11	1	131
No. of Bidders	1,561	6.50	6	3.91	1	25

Note. Sample: 1,561 completed auctions with certified grades, after dropping private listings and incomplete bid histories. Book value differs in N by two auctions for which NGC reports no published value at the relevant grade level.

Table 2 reports pairwise correlations among auction-level variables. Winning bids correlate positively but weakly with both bid count (0.28) and bidder count (0.19); the strong correlation between bid count and bidder count (0.82) reflects the mechanical proxy-bidding relationship rather than economic content. The modest correlation between winning bids and participation measures motivates including $\log(n_{\text{bidders}})$ as a competition control in one of the specifications in Section 4.

Table 2: Correlation Matrix for Auction-Level Variables

	Winning Bid	No. of Bids	No. of Bidders
Winning Bid	1.0000		
No. of Bids	0.2840	1.0000	
No. of Bidders	0.1943	0.8194	1.0000

3.2 Construction of Behavioural Metrics

Detection Objective and Approach

The objective of the detection step is to develop empirical markers of shill bidding visible in observed bid data, then look for them. Each of the three shilling types reviewed in Section 2 predicts distinct observable behaviour, and the markers are designed to capture those predictions:

- *Signal-effect* shilling requires early bids to fake participation and interest, producing systematically early bidding patterns.
- *Price-squeezing* shilling exploits eBay’s proxy mechanism through incremental probing of the winner’s maximum, producing long sequences of small successive bids.
- *Cross-cutting* (relational) features apply across types: shilling requires the bidder to be aligned with the seller (whether the seller bidding through an alternate account, or

a confederate), so suspicious bidders concentrate participation with specific sellers, and they must avoid winning auctions to preserve the seller’s revenue.

The three types differ in their detectability from platform-outsider data, however: as discussed in Section 2, only price-squeezing leaves a trace visible in public bid histories.

A complete structural model of bidder timing is not feasible in this setting, and existing online-auction models do not address shill bidding directly: they either restrict each bidder to a single bid (Bajari & Hortaçsu, 2003), abstract from within-auction sequential bidding (Hickman, 2010; Hickman et al., 2017), or model cross-auction dynamics without within-auction sequence (Hickman et al., 2021). I therefore provide economic intuition for each metric and ground each in a theoretical prediction from the shill bidding literature, rather than deriving the metrics from a full structural model.

The metrics deliberately exclude auction prices. Using price information to construct detection features would create mechanical correlation when prices subsequently serve as the dependent variable in causal analysis. Price independence also enables real-time detection before auctions conclude.

The computer-science fraud-detection literature has applied hierarchical clustering on hand-engineered behavioural features to eBay-style auction data (Alzahrani & Sadaoui, 2018, 2020; Anowar et al., 2018). The set of metrics I use largely overlaps with the markers introduced by Alzahrani and Sadaoui (2018): *Seller Loyalty* (their *Bidder Tendency*), *Early Bidding*, and *Winning Ratio* are identical in formulation; *Mean Bid Time* aggregates their bidder-auction *Early Bidding* metric to the bidder level; *Successive Proxy Outbidding* refines their binary $\{0, 0.5, 1\}$ *Successive Outbidding* indicator into a sequence-length measure that explicitly captures the proxy-mechanism dynamics described by Engelberg and Williams (2009); and *Bidding Ratio* re-formulates their within-auction bid-share metric as a cross-auction measure. Theoretical motivation for the marker set draws on Marshall (2024), whence the identification framework for English auctions with shill bidding, and on Engelberg and Williams (2009) for the proxy-bidding mechanism.

I construct six behavioural metrics at the bidder-auction level, meaning each metric characterises how a specific bidder behaved in a specific auction. This level of analysis preserves richer variation than aggregating to either the bidder level or auction level alone. All metrics are scaled to $[0, 1]$ by min-max normalisation within the sample, with higher values indicating more suspicious behaviour. Rank ordering is preserved, and the scaling prevents any single metric’s native range from dominating the distance calculations in the clustering step.

Markers by Shilling Type

Table 3 groups the six metrics by the type of shilling each predicts.

Table 3: Behavioural markers grouped by shilling type

Shilling Type	Predicted Behaviour	Markers
Signal-effect	Early bids to fake interest	Early Bidding, Mean Bid Time
Price-squeezing	Many incremental bids per auction	Successive Proxy Outbidding, Bidding Ratio
Across types (relational)	Repeat seller; avoid winning	Seller Loyalty, Winning Ratio

Two consequences for the clustering: each shilling type produces a distinct behavioural signature, so the detection step should not search for a single homogeneous “shill cluster” but for a multimodal structure in which different combinations of elevated metrics correspond to different strategies; and the cross-cutting markers (Seller Loyalty, Winning Ratio) should be elevated in any cluster of genuine shills regardless of strategy, since both follow from the underlying seller-bidder coordination rather than from any specific bidding tactic.

The detectability asymmetry has direct empirical consequences. Reserve-price shilling is unobservable in this dataset because the reserve threshold is private. Signal-effect bidding produces a behavioural pattern that is difficult to distinguish from honest enthusiast early bidding, since legitimate collectors in this category routinely bid early as well. Price-squeezing, in contrast, leaves a mechanically deterministic signature through eBay’s proxy mechanism that has no common honest analogue. I therefore expect the recovered suspicious cluster to be dominated by price-squeezing and cross-cutting coordination markers, with timing-based signal-effect markers showing weaker separation; Section 4 confirms this pattern, with elevated Successive Proxy Outbidding and Seller Loyalty driving the cluster identification and the timing markers contributing comparatively little.

Bidder-Auction Level Metrics

Seller Loyalty (SL). *Type: across types (relational). Prior usage: identical formulation to Alzahrani and Sadaoui 2018, where it is called Bidder Tendency (BT); theoretical interpretation Marshall 2024.* Shill bidders systematically participate in auctions run by their accomplice seller rather than bidding randomly across the platform. For bidder i and seller s , I define:

$$SL(i, s) = \frac{\text{auctions bid by } i \text{ for seller } s}{\text{total auctions bid by } i}$$

I refer to this metric as “Seller Loyalty” rather than the original “Bidder Tendency” (Alzahrani & Sadaoui, 2018) inasmuch as the former more accurately describes what the

metric captures: a bidder’s tendency to participate with the same seller. Henceforth I use the abbreviation SL, and reserve BT for references to the prior literature. Honest bidders search for items based on characteristics rather than seller identity, producing lower and more dispersed values. High Seller Loyalty reveals the bidder-seller concentration characteristic of shilling, where the bidder operates on behalf of the seller. Bidders who participate in only a single auction are assigned $SL = 0$, since a single observation provides no information about seller concentration and the mechanical value of 1 would be a normalisation artefact.

Early Bidding (EB). *Type: signal-effect. Prior usage: identical to Alzahrani and Sadaoui 2018; theoretical interpretation Marshall 2024.* Signal-effect shilling requires early bidding to create the appearance of interest and demand. For bidder i in auction a with duration D_a , first bid time $t_{i,a}$, and auction start time t_a^{start} :

$$EB(i, a) = 1 - \frac{t_{i,a} - t_a^{start}}{D_a}$$

Values near 1 indicate very early bidding. Honest bidders frequently employ late bidding strategies to avoid revealing information (Ockenfels & Roth, 2006; Roth & Ockenfels, 2002), whilst shills bid early to maximise visibility and signal effect (Lamy, 2009).

Mean Bid Time (MBT). *Type: signal-effect. Prior usage: aggregates the bidder-auction EB metric of Alzahrani and Sadaoui 2018 to the bidder level.* Aggregating early bidding behaviour across all auctions in which bidder i participated reveals persistent timing patterns:

$$MBT(i) = \text{mean}(EB(i, a) \mid i \text{ bid in } a)$$

Consistently high MBT values suggest systematic rather than auction-specific early bidding, characteristic of strategic manipulation rather than idiosyncratic bidder preferences.

Successive Proxy Outbidding (SPO). *Type: price-squeezing. Prior usage: refines the binary $\{0, 0.5, 1\}$ Successive Outbidding metric of Alzahrani and Sadaoui 2018; theoretical interpretation from Engelberg and Williams 2009 and Marshall 2024.* Price-squeezing shills exploit eBay’s proxy bidding mechanism whereby incremental bids probe the current high bidder’s maximum without winning. Each shill bid prompts an automatic proxy outbid by the high bidder; the shill increments to the next standard increment and prompts another proxy outbid. The length of this successive proxy-outbidding sequence, ending when the bid increment changes (signalling the shill stopped just short of the opponent’s maximum), captures the discover-and-stop strategy whence the metric draws its motivation.

Formally, I count, for each bidder i , the longest run of consecutive manual bids by i

each immediately followed by an automatic proxy bid from a different bidder. The run terminates when i stops bidding or places a bid not followed by a proxy outbid (because the opponent’s proxy maximum has been exceeded). Table 4 walks through a stylised example with the same structure as the real auction reproduced in Table 8 below. Bidder H has set a hidden proxy maximum of \$9.40. S places successive manual bids below that maximum, and each prompts an automatic proxy outbid by H . At row 11, S ’s bid of \$9.20 elicits a proxy response of only \$9.40, a gap of \$0.20 rather than the standard \$0.50, because the proxy can only respond up to H ’s ceiling. This gap compression reveals H ’s maximum to S ; under the discover-and-stop strategy, S stops bidding here, ending the run at length 5.

Table 4: Worked example of Successive Proxy Outbidding (run length 5)

#	Bidder	Type	Amount	Gap	Note
1	H	(proxy max set)	\$9.40	–	H ’s maximum, hidden from S
2	S	Manual	\$5.50	–	initial probe
3	H	Auto	\$6.00	\$0.50	proxy outbid (run = 1)
4	S	Manual	\$6.50	–	
5	H	Auto	\$7.00	\$0.50	proxy outbid (run = 2)
6	S	Manual	\$7.50	–	
7	H	Auto	\$8.00	\$0.50	proxy outbid (run = 3)
8	S	Manual	\$8.50	–	
9	H	Auto	\$9.00	\$0.50	proxy outbid (run = 4)
10	S	Manual	\$9.20	–	
11	H	Auto	\$9.40	\$0.20	proxy outbid (run = 5); gap compresses; S stops

Note. Stylised example. “Manual” is a bid placed by the suspected shill S ; “Auto” is an automatic proxy response from H up to H ’s hidden ceiling of \$9.40. “Gap” is the price increase from the prior automatic response, normally one standard increment (\$0.50 in this band). The bold row marks the point at which the gap compresses below the standard increment, signalling that H ’s ceiling has been reached and triggering S to stop.

Long runs indicate systematic probing rather than honest bidding, which typically involves fewer, larger jumps closer to the bidder’s true valuation. Gap compression at the end of a run, although not part of the metric itself, signals that the opponent’s proxy ceiling has been reached, the empirical trigger that, under the discover-and-stop strategy, causes the shill to stop probing.

Winning Ratio (WR). *Type: across types (relational).* Prior usage: identical to Alzahrani and Sadaoui 2018; theoretical interpretation Chakraborty and Kosmopoulou 2004. Shills must avoid winning auctions to prevent paying the seller. For bidder i participating in N_i auctions:

$$WR(i) = 1 - \frac{\text{auctions won by } i}{\text{auctions where } i \text{ placed } \geq 2 \text{ bids}}$$

High winning ratios indicate frequent participation without victories, suggesting bids placed to influence rather than purchase. The denominator restriction to auctions with multiple bids excludes accidental single exploratory bids.

Bidding Ratio (BR). *Type: price-squeezing. Prior usage: re-formulates the within-auction bid-share metric of Alzahrani and Sadaoui 2018 as a cross-auction measure.* Frequent participation to raise prices incrementally characterises some shill strategies. I compute the ratio of auctions where bidder i bid multiple times to total participation, capturing intensity of bidding activity. Whereas their formulation defines BR as bidder i 's share of bids within a single auction, the cross-auction formulation here captures persistent multi-bid behaviour across i 's participation history, a marker of repeated price-squeezing rather than one-off active bidding.

3.3 CURE Clustering for Shill Detection

Clustering Approach Rationale

Shill bidding detection faces a labelling problem: eBay does not publicly disclose confirmed fraud cases, so the data carry no ground-truth indicator. Supervised learning is therefore unavailable, and the natural alternative is unsupervised clustering.

Among the unsupervised options, I prefer hierarchical methods to partitioning approaches like k-means: they do not require specifying cluster shapes a priori (allowing non-spherical clusters), and the dendrogram exposes cluster relationships at multiple scales. Alzahrani and Sadaoui (2020) report that hierarchical methods outperform partitioning ones on auction-fraud data; within the hierarchical family they find the Clustering Using REpresentatives (CURE) algorithm (Guha et al., 1998) to dominate BIRCH and single-linkage on the same benchmarks. I adopt CURE on that basis.

CURE Algorithm Implementation

CURE departs from classical hierarchical clustering in two respects. Rather than representing each cluster by a single centroid or by all its constituent points, CURE picks c well-scattered representative points (RPs) that capture the cluster's shape and extent; I set $c = 4$, enough to capture non-spherical geometries without the computational cost of large c . Inter-cluster distance is the minimum distance between any pair of RPs drawn from the two clusters.

A shrinking factor $\alpha \in [0, 1]$ then pulls each RP towards the cluster centroid, dampening outlier influence. $\alpha = 0$ recovers minimum spanning tree clustering (very outlier-

sensitive); $\alpha = 1$ recovers centroid clustering (no shape flexibility). I set $\alpha = 0.0001$, the value used by Alzahrani and Sadaoui (2020) on a closely related dataset.

The algorithm proceeds iteratively, starting with each observation as its own cluster. At each step, the two clusters with minimum inter-RP distance merge, whereupon new representative points are selected from the merged cluster and shrunk towards the new centroid by factor α . The process continues until reaching a pre-specified number of clusters determined by the silhouette method described below. Distance calculations exploit a k-d tree data structure for computational efficiency.

The worst-case time complexity of CURE is $O(N^2 \log N)$ where N is the sample size (Guha et al., 1998). While substantial for large N , this complexity is acceptable because clustering is performed offline as a one-time operation. My implementation uses a Python adaptation of the original CURE algorithm.

Determining Optimal Number of Clusters

The number of clusters K has to be chosen ex ante. Too few and the suspicious behaviour blurs into the normal; too many and the partition fragments into noise. I choose K by the silhouette method (Rousseeuw, 1987), whereby cluster quality is measured by both the tightness of observations within clusters and the separation between clusters.

For each observation, the silhouette coefficient compares the average distance to other observations in the same cluster versus the average distance to observations in the nearest other cluster. Values near 1 indicate tight, well-separated clusters, while values near 0 or negative suggest poor cluster assignment. The overall silhouette score averages individual coefficients across all observations.

I apply k-means clustering for 2 through 20 clusters and compute the average silhouette score for each K . The optimal number of clusters maximises the silhouette score. This procedure yields an optimal cluster count of 12 clusters. I then run CURE with this cluster count using the representative point and shrinking factor parameters specified above. Because MBT is the cross-auction mean of EB, the two metrics are mechanically related; I scale each by $\sqrt{0.5}$ prior to computing distances so that the early-bidding construct receives the same total weight as any single metric.

Labelling Clusters as Suspicious or Normal

After CURE partitions the data into clusters, I must classify clusters as representing either normal or suspicious bidding behaviour. Following Alzahrani and Sadaoui (2020), I adopt a threshold-based labelling rule: for each cluster, compute the mean value of each metric

across all observations in that cluster, then average these means across metrics to obtain an overall cluster suspiciousness score. If this score exceeds $\mu + 0.5\sigma$, where μ and σ are the mean and standard deviation of metric values in the full sample, the cluster is labelled suspicious; otherwise it is labelled normal. The labelling is robust to nearby threshold values: Cluster 12 is uniquely identified for any $\mu + x\sigma$ rule with $0.40 < x < 0.62$, placing the chosen $x = 0.5$ comfortably in the middle of the stable interval.

This threshold balances false positives (normal bidders flagged as suspicious) against false negatives (bona fide shills missed). The $\mu + 0.5\sigma$ cutoff follows Alzahrani and Sadaoui (2020)’s experiments and reflects the fact that shill bidders are a small minority of the sample with distinctively elevated metric values; balanced-cluster-size criteria would tend to miss them, so a labelling rule is the standard remedy. All bidder-auction pairs assigned to suspicious clusters receive a binary “shill” label of 1; those in normal clusters receive label 0.

An auction is classified as suspicious if it contains any bidder-auction pairs labelled suspicious. This auction-level classification serves as the treatment variable in subsequent causal analysis. The final labelled dataset contains 144 suspicious auctions (9.2%) and 1,417 normal auctions (90.8%).

3.4 OLS Specification

Having labelled auctions as flagged or unflagged, the main empirical exercise is to estimate the effect of shill bidding on the sale-to-book-value ratio. The challenge is selection: sellers may shill selectively on items with particular characteristics (high-value items, particular grades, auctions with few initial bidders), so a naive comparison of flagged and unflagged auction outcomes confounds the treatment effect with selection on those characteristics. Section 4.5 develops the identification strategy that selects among the estimators introduced in this and the following subsection.

The baseline OLS specification is:

$$\frac{\text{SalePrice}_i}{\text{BookValue}_i} = \beta_0 + \beta_1 \times \mathbb{1}(\text{Shill}_i) + \beta_2 \times \text{GradeFE}_i + \epsilon_i$$

where $\mathbb{1}(\text{Shill}_i)$ is the binary indicator for detected shill bidding, and GradeFE_i represents grade fixed effects. Standard errors are clustered by country of origin to account for within-country correlation of auction outcomes (130 clusters), following the practical recommendations in MacKinnon, Nielsen, and Webb (2023). The coefficient β_1 provides an estimate of the treatment effect under the assumption that conditional on grade fixed

effects, shill bidding is independent of unobserved determinants of the sale price to book value ratio.

3.5 Propensity Score Matching

Motivation for PSM

Propensity score matching builds a comparison group of unflagged auctions resembling the flagged ones on observables, and recovers the ATT under Rosenbaum and Rubin (1983)'s conditional-independence assumption.

Treatment Definition and Sample

The treatment is defined as a binary indicator equal to 1 if the auction contains one or more bidder-auction pairs classified as suspicious by the CURE clustering procedure, and 0 otherwise. The treated group consists of 144 suspicious auctions; the control pool contains 1,417 normal auctions. All auctions with valid book value data are included in the matching analysis.

Propensity Score Estimation

I estimate propensity scores using a logit model where the dependent variable is the binary treatment indicator and covariates include:

- Grade fixed effects: Binary indicators for each certified grade level, controlling for item quality heterogeneity
- Book value: The NGC price guide value in dollars, controlling for item value

These covariates capture observable characteristics that may simultaneously affect both the probability of shill bidding and auction outcomes. For instance, high-value items may attract both more shill activity (the potential gains are larger) and higher prices (the intrinsic value is higher). Including book value as a covariate isolates the treatment effect from this confound. I do not include the number of bidders or bids as covariates since these are plausibly affected by shill bidding itself and would constitute bad controls.

After estimating the propensity score model, I assess covariate balance before and after matching using standardised bias statistics and t-tests for mean differences between treated and control groups.

Matching Implementation

I implement two complementary matching strategies. First, propensity score matching estimates the ATT by matching treated auctions to control auctions with similar estimated propensity scores based on grade and book value. Secondly, nearest-neighbour matching with exact matching on coin type (defined by country, catalogue number, and grade) provides a more conservative estimate by ensuring treated and control auctions involve the same type of coin, at the cost of a smaller matched sample. Matched coins may still differ in other characteristics such as year, mintmark, or condition details beyond the certified grade. The exact matching requirement drops treated observations without any control auction for the same coin type, reducing the sample from 1,559 to 791 observations but strengthening the comparability of matched pairs.

Treatment Effect Estimation

The estimand is the average treatment effect on the treated:

$$\tau_{ATT} = E \left[\frac{\text{SalePrice}_i}{\text{BookValue}_i} \mid \text{Shill}_i = 1 \right] - E \left[\frac{\text{SalePrice}_i}{\text{BookValue}_i} \mid \text{Shill}_i = 0, \text{Matched} \right].$$

For propensity score matching, I report Abadie and Imbens (2006) robust standard errors that account for estimation uncertainty in the propensity score. For nearest-neighbour matching, I report standard errors under the assumption of independent observations within matched pairs.

4 Results

4.1 CURE Clustering Results

I run CURE on the 10,151 bidder-auction pairs across the 1,561 auctions, with the six metrics from Section 3.2 as input.

Optimal Cluster Selection

Sweeping K from 2 to 20, the silhouette score peaks at $K = 12$. I run CURE with this cluster count and the parameters fixed in Section 3.3 ($c = 4$, $\alpha = 0.0001$).

Cluster Composition

Table 5 reports the resulting cluster sizes, ranging from 144 (Cluster 12) to 2,548 (Cluster 8). The spread reflects real heterogeneity in bidding behaviour: profiles such as late bidding with light participation are far more common than the tight, repetitive patterns that fill the smaller clusters.

Table 5: Cluster Sizes from CURE Algorithm

Cluster	1	2	3	4	5	6	7	8	9	10	11	12
Size	261	1,399	581	675	1,184	450	796	2,548	1,397	515	201	144

Classification of Suspicious Cluster

Applying the labelling rule from Section 3.3, Cluster 12 emerges as the sole suspicious cluster: 144 pairs (1.4% of the total), with the remaining 10,007 pairs distributed across the other eleven.

At the auction level, an auction is flagged if it contains any Cluster 12 pair. This yields 144 suspicious auctions (9.2%) and 1,417 normal auctions (90.8%). In this sample each flagged auction happens to contain exactly one Cluster 12 pair, so the bidder-pair and auction-level flag rates coincide.

Absent ground truth, the 144 flagged pairs are best read as an upper bound on the incidence of price-squeezing shill bidding among these bidder–auction pairs, since sincere strategies can in principle reproduce its markers, as when a bidder manually increments against the standing proxy rather than delegating to the proxy system. They are not a bound on shill bidding overall: reserve-price shilling is unobservable in these data, and signal-effect shilling does not separate from honest early bidding (Section 2), so total incidence may lie on either side of 9.2 per cent.

4.2 Behavioural Characteristics of Suspicious Cluster

Table 6 compares the mean values of the six behavioural metrics between Cluster 12 (the suspicious cluster) and the pooled normal clusters. The table reports means and standard deviations for each metric, scaled to the [0,1] interval with higher values indicating more suspicious behaviour.

Cluster 12 exhibits elevated values on every metric the theory associates with shill bidding (Table 6). The largest gap is in Successive Proxy Outbidding, two orders of magnitude above the normal-cluster mean and consistent with the price-squeezing trace

Table 6: Behavioural Metrics: Cluster 12 vs. Normal Clusters

Metric	Cluster 12 (Susp.)		Normal Clusters		Diff	p
	Mean	SD	Mean	SD		
Seller Loyalty (SL)	0.580	0.130	0.226	0.183	+0.354	$< 10^{-69}$
Early Bidding (EB)	0.701	0.157	0.572	0.389	+0.129	$< 10^{-16}$
Mean Bid Time (MBT)	0.632	0.082	0.573	0.132	+0.059	$< 10^{-13}$
Successive Proxy Outbid (SPO)	0.673	0.092	0.007	0.033	+0.665	$< 10^{-125}$
Winning Ratio (WR)	0.911	0.118	0.760	0.277	+0.151	$< 10^{-31}$
Bidding Ratio (BR)	0.604	0.135	0.356	0.163	+0.248	$< 10^{-47}$

Note. Welch’s t -test (unequal variances). Cluster 12: $n = 144$ pairs. Normal clusters: $n = 10,007$ pairs.

described by Engelberg and Williams (2009). The cross-cutting coordination markers Seller Loyalty and Winning Ratio are also markedly elevated: the former roughly twice the normal level, the latter approaching one (so suspicious bidders rarely win despite heavy participation), in line with the bidder–seller coordination characterised by Marshall (2024). Bidding Ratio is elevated by roughly 70%.

The timing metrics show smaller gaps than the price-squeezing and cross-cutting markers, though Mean Bid Time still differs meaningfully (Cluster 12 mean 0.63 vs. normal 0.57) and Early Bidding less so (0.70 vs. 0.57, the largest normal-cluster variance among the six). As discussed in Section 2, early bids by coin enthusiasts are hard to distinguish from early shill bids on timing alone, so the clustering recovers primarily the price-squeezing and cross-cutting coordination signature.

Individual Metric Classification Performance

To characterise each metric’s contribution further, Table 7 evaluates each metric as a standalone binary classifier. For each metric, I sweep thresholds and report the one that maximises the F1 score, F1 being the harmonic mean of precision (the share of flagged pairs that are truly suspicious) and recall (also called sensitivity: the share of truly suspicious pairs that are flagged). The table also reports the false negative and false positive counts at that threshold, the corresponding specificity (the share of truly normal pairs correctly classified), and the area under the receiver operating characteristic curve (AUC), a standard threshold-independent measure of classification quality (Fawcett, 2006).

“Ground truth” for these classification metrics is the binary label assigned by the clustering, since no externally validated ground truth is available; the metrics therefore characterise how well each feature reproduces the multivariate clustering boundary, not necessarily how well it identifies bona fide shill bidding.

Table 7: Individual Metric Classification Performance (F1-Optimal Threshold)

Metric	Threshold	FN	FP	Sensitivity	Specificity	AUC
SPO	0.571	0	0	100.0%	100.0%	1.000
SL	0.625	61	278	57.6%	97.2%	0.929
BR	0.625	50	530	65.3%	94.7%	0.872
WR	0.952	54	2,978	62.5%	70.2%	0.673
MBT	0.670	90	1,632	37.5%	83.7%	0.671
EB	0.492	2	6,004	98.6%	40.0%	0.534

Note. “Ground truth” is the cluster label assigned by CURE, since no externally validated labels exist. For each metric I sweep thresholds and report the one maximising F1 (the harmonic mean of precision and recall). Sensitivity is recall on flagged pairs; specificity is the share of normal pairs correctly classified; AUC is the area under the ROC curve. Columns characterise how well a single metric reproduces the multivariate clustering boundary, not necessarily how well it identifies bona fide shill bidding.

SPO achieves perfect separation between Cluster 12 and the normal clusters: at a threshold of 0.571, every suspicious pair is correctly classified and no normal pair is misclassified, yielding an AUC of 1.000. This is an ex post finding, not an ex ante assumption (the clustering operates in the full six-dimensional metric space without privileging any single metric), but it carries a substantive implication: successive proxy outbidding is the defining behavioural signature of the suspicious cluster.

The other metrics contribute to determining the cluster boundary. Some observations have moderately elevated SPO scores but are assigned to normal clusters; these borderline cases have low Seller Loyalty and Bidding Ratio, lacking the seller-specific concentration and repeated bidding that characterise the suspicious cluster. The clustering algorithm distinguishes these cases from genuine shill activity by evaluating the full behavioural profile rather than SPO alone. A simple SPO threshold would either include these false positives or require an arbitrary cutoff; multivariate clustering finds the boundary where elevated SPO co-occurs with the broader behavioural pattern predicted by shill bidding theory.

The remaining metrics exhibit varying degrees of standalone discriminatory power. SL and BR achieve high AUC values (0.929 and 0.872 respectively) but face a sensitivity–specificity tradeoff at their F1-optimal thresholds, correctly identifying 58–65% of suspicious pairs while maintaining specificity above 94%. WR and MBT achieve moderate AUCs (0.673 and 0.671 respectively) on different tradeoffs: WR labels with reasonable sensitivity at the cost of many false positives, MBT vice versa. Early Bidding alone performs near chance (AUC 0.534), confirming that the within-auction timing component contributes essentially no separation on its own; the modest MBT signal reflects the bidder-level ag-

gregation rather than within-auction timing. SPO is therefore the primary discriminator, but SL and BR sharpen the cluster boundary by excluding borderline-SPO cases whose wider behavioural profile is inconsistent with coordinated shilling.

4.3 Anatomy of a Shill Auction

To illustrate the bidding pattern that the SPO metric captures, I examine one of the 144 flagged bidder-auction pairs in Cluster 12. The auction is for a USA 10c 1917 (KM 140) graded MS-65, with a book value of \$375 and a final sale price of \$306.20. Over the course of the auction, the suspected shill bidder B0625 placed nine manual bids; the seven that form the longest consecutive proxy-outbid run are reproduced in Table 8, giving an SPO_{raw} score of 7.

Each of B0625’s manual bids in the run is immediately followed (within seconds, or in pair 1’s case within a few minutes) by an automatic proxy response from bidder B0746. The gap between each shill bid and the proxy response holds steady at the standard eBay increment: \$2.50 for the first three pairs and \$5.00 for pairs four through six (the increment increases at the \$250 boundary). At row 14, the gap compresses to \$1.84: the proxy system can only respond up to B0746’s maximum of \$281.84, which falls short of the next standard increment (\$285.00). The compression itself does not enter the SPO calculation, but it signals to the shill that B0746’s ceiling has been located, and under the discover-and-stop strategy, this is the empirical trigger to stop bidding. The shill’s withdrawal is what ends the run; further probing would risk winning the auction outright.

Table 8: SPO probing run: USA 10c 1917 (KM 140), MS-65

#	Bidder	Type	Amount	Gap	Time
1	B0625	Manual	\$237.50		Aug 31 8:44p EDT
2	B0746	Auto	\$240.00	\$2.50	Aug 31 8:53p
3	B0625	Manual	\$242.50		Aug 31 10:26p
4	B0746	Auto	\$245.00	\$2.50	Aug 31 10:26p
5	B0625	Manual	\$245.00		Aug 31 11:53p
6	B0746	Auto	\$247.50	\$2.50	Aug 31 11:53p
7	B0625	Manual	\$250.00		Sep 1 1:32a
8	B0746	Auto	\$255.00	\$5.00	Sep 1 1:32a
9	B0625	Manual	\$260.00		Sep 1 3:02a
10	B0746	Auto	\$265.00	\$5.00	Sep 1 3:02a
11	B0625	Manual	\$270.00		Sep 1 4:43a
12	B0746	Auto	\$275.00	\$5.00	Sep 1 4:43a
13	B0625	Manual	\$280.00		Sep 1 6:19a
14	B0746	Auto	\$281.84	\$1.84	Sep 1 6:19a

The shill withdraws after row 14, never becoming the winning bidder. Bidder B0746

later returns with a higher maximum proxy and ultimately wins the auction at \$306.20. This pattern (incremental probing, gap compression at the opponent’s proxy ceiling, withdrawal before winning) is representative across the 144 flagged auctions and is the signal that the SPO metric is designed to detect.

4.4 Auction-Level Descriptive Comparison

Table 9 presents summary statistics comparing auction outcomes for the 144 suspicious auctions (those containing at least one Cluster 12 bidder-auction pair) against the 1,417 normal auctions. The comparison covers item value (book value), the normalised price outcome (sale-to-book ratio), and the bidding-intensity variables (winning bid, number of bids, number of bidders).

Table 9: Summary of Auctions by Classification

Class	Variable	N	Mean	Median	Std. Dev.	Min	Max
Normal	Book Value (\$)	1,415	163.06	79.00	397.52	0.30	5,079.22
	Sale / Book Ratio	1,415	0.843	0.699	1.666	0.010	50.000
	Winning Bid (\$)	1,417	109.69	50.00	281.64	0.45	3,605.55
	No. of Bids	1,417	21.94	19	17.59	1	131
	No. of Bidders	1,417	6.55	6	4.06	1	25
Suspicious	Book Value (\$)	144	183.31	90.00	281.23	1.50	2,500.00
	Sale / Book Ratio	144	1.008	0.906	0.427	0.117	2.300
	Winning Bid (\$)	144	158.06	79.50	221.18	2.28	1,634.00
	No. of Bids	144	30.37	29	8.13	16	60
	No. of Bidders	144	6.05	6	1.92	3	12

Note. Welch’s t -test on mean differences (Suspicious – Normal): Book Value +\$20.25 ($t = 0.79$, $p = 0.43$); Sale/Book Ratio +0.165 ($t = 2.90$, $p = 0.004$); Winning Bid +\$48.37 ($t = 2.43$, $p = 0.016$); No. of Bids +8.42 ($t = 10.23$, $p < 10^{-20}$); No. of Bidders -0.50 ($t = -2.59$, $p = 0.010$). Book Value and Sale/Book Ratio drop two normal auctions for which NGC reports no published value at the relevant grade.

Flagged auctions sell for roughly 44% more at the mean raw winning bid (\$158 vs. \$110), but this confounds two channels. Item value is statistically indistinguishable across classes: flagged book values are slightly higher on average (\$183 vs. \$163) but the difference is not statistically significant ($p = 0.43$). The substantive comparison is the sale-to-book ratio, which is 16.5 percentage points higher in flagged auctions ($p = 0.004$). Flagged auctions also attract about a third more bids (30 vs. 22) without attracting more unique bidders (6.1 vs. 6.5), so the extra bid volume reflects iterative price-probing rather than additional bidder entry (Lamy, 2009). The regression and matching estimates that follow control for item heterogeneity and seller heterogeneity to isolate the causal effect.

4.5 Identification Strategy

The object of interest is the average treatment effect on the treated (ATT): the price premium that flagged auctions earn relative to what those same auctions would have sold for absent shilling. Detection does not deliver this. Because the seller is the one shilling, treatment is not assigned randomly across the platform, and selection enters at two levels.

Between-seller selection. The bulk of the 144 suspicious auctions come from 35 sellers (of the 172 in the sample) who also run unflagged auctions; the sellers who shill differ from those who do not on many dimensions. Sellers who run successful listings (good photography, fast shipping, careful descriptions) may also be the sellers who shill, in which case a comparison of flagged to unflagged auctions in the full sample partly reflects seller quality rather than the shill effect itself.

Table 10 characterises the 35 mixed-portfolio sellers. The distribution is sharply right-skewed. The median seller runs only 5 auctions (1 flagged); the top seller runs 348 (41 flagged). The flagged share within a seller’s portfolio ranges from 1.2% to 66.7%, with a median of 20%. This heterogeneity matters for two reasons. First, the within-seller specification is identified off variation in flagged status *within* seller, so sellers with very small portfolios (a single flagged auction and one or two non-flagged) contribute relatively little leverage; the bulk of identifying variation comes from a smaller number of high-volume sellers. Secondly, the wide range of flagged shares limits any concern that the 35 are a homogeneous subpopulation of habitual shillers; some shill in a small fraction of their portfolios, others in two-thirds.

Table 10: Distribution across the 35 mixed-portfolio sellers

Per-seller variable	Mean	Median	Std. Dev.	Min	P25	P75	Max
Total auctions	35.2	5	72.6	2	3	31.5	348
Flagged auctions	3.8	1	7.6	1	1	2	41
Non-flagged auctions	31.4	4	65.6	1	2	28	307
Flagged share (%)	25.1	20.0	19.0	1.2	9.4	33.3	66.7

Note. Sample: the 35 sellers whose portfolios contain both flagged and non-flagged auctions (1,232 total auctions, of which 132 flagged and 1,100 non-flagged; 1,231 with a non-missing book value enter the within-seller regression). Each row reports the distribution across sellers of the per-seller variable. “Flagged share” is the share of a given seller’s auctions classified as suspicious by the clustering procedure.

Within-seller selection. A given seller’s choice of which auctions in the portfolio to shill is also unlikely to be random. Plausible motives include shilling auctions where the seller anticipates thin organic bidding, where the seller’s hidden reserve exceeds the likely market clearing price, or where the item’s book value most understates the seller’s private valuation. The direction of bias depends on whether the auctions chosen for shilling

would otherwise have outperformed or underperformed the rest of the seller’s portfolio, and is not signed by theory alone.

I address the two layers with complementary designs. The within-seller specification handles between-seller selection: I restrict the sample to the 35 mixed-portfolio sellers, include seller fixed effects, and identify the skill premium from variation within seller, differencing out any time-invariant seller characteristic. The auction-level layer is addressed in two further steps. First, exact-match nearest-neighbour designs at the seller-coin-grade level hold item identity nearly fixed: each flagged auction is matched to a non-flagged auction of the same coin at the same grade run by the same seller, leaving little room for selection on observable item characteristics within a seller’s portfolio. Secondly, Oster (2019) bounds quantify how strong unobservable auction-level selection would have to be to nullify the within-seller estimate.

The descriptive evidence on within-seller observable selection is reassuring. For each of the 35 mixed-portfolio sellers, I compute the within-seller mean difference (flagged minus non-flagged) on book value, grade, and year of issue, then average these per-seller differences across the 35 sellers (Table 11).¹ None of the within-seller differences is statistically distinguishable from zero. Book values differ by \$45 on average (SE \$44, $p = 0.32$), grades by 1.2 points (SE 1.4, $p = 0.40$), and years of issue by 12 years (SE 7.5, $p = 0.12$). One observable does differ within seller: country composition (a chi-square test across 97 origins yields $p = 0.012$), reflecting that some sellers concentrate shilling on particular national segments of the catalogue. The exact-match design at the seller-country-grade-KM level absorbs this directly. Under Oster’s proportional-selection assumption, the small observable selection on item characteristics bounds unobservable within-seller selection along the same dimensions, though that assumption requires care here (Section 4.6).

Together the exact-match designs, Oster bounds, and observable-comparison evidence address most of the within-seller auction-level selection that can be addressed without ground truth; none rules out residual selection on unobservables that cannot be proxied by observables.

A separate concern is structural correlation between the detection metrics and the outcome. Two of the six markers (Successive Proxy Outbidding and Winning Ratio) are computed from the same bid sequences that determine the sale price, so mechanical

¹The estimator weights each seller equally, regardless of portfolio size. A pooled within-seller OLS regression of each observable on the flag indicator with seller fixed effects (weighting sellers by $N_s p_s (1-p_s)$) yields qualitatively similar conclusions on each observable: book value +\$17 (SE \$13, $p = 0.18$), grade +0.97 (SE 0.69, $p = 0.16$), year of issue -4.16 (SE 3.10, $p = 0.18$). The per-seller estimator reported in the table is the more conservative of the two.

Table 11: Within-seller observable comparison: average across 35 mixed-portfolio sellers

Variable	Avg. per-seller diff. (flagged – non-flagged)	SE	p	N sellers
Book value (\$)	+45.06	44.37	0.32	35
Grade	+1.23	1.44	0.40	33
Year of issue	–11.93	7.55	0.12	35

Note. For each seller s in the 35 mixed-portfolio sellers, the within-seller difference is $\bar{x}_{s,\text{flagged}} - \bar{x}_{s,\text{non-flagged}}$. The table reports the simple average of these differences across sellers and the SE computed across sellers (i.e., $SD/\sqrt{G}$ with G the seller count). The grade row uses $G = 33$ because two sellers have no flagged auctions with a non-missing numerical grade. Country composition differs within seller (a chi-square test across 97 origins yields $p = 0.012$) and is absorbed by the exact-match designs reported in Table 12.

correlation between flag and outcome could exist independently of any treatment effect. The remaining four metrics are price-sequence-independent, but since SPO alone perfectly separates the suspicious cluster from the normal clusters (Section 4.2), dropping it from the clustering input would amount to redefining the treatment rather than testing its robustness.

There is also classification noise to consider: the treatment indicator is the output of an unsupervised clustering procedure, not ground truth, so misclassification attenuates the estimate towards zero rather than away from it. The Oster (2019) bounds reported below speak to the unobservable-selection version of this threat, with the caveat discussed in Section 4.6.

4.6 Estimates

The outcome is the ratio of sale price to NGC book value. Normalising by book value allows comparison across coins of different intrinsic worth: a coefficient of 0.15 means flagged auctions sell for 15% more relative to book value, regardless of whether the coin is worth \$50 or \$500. Book values are constructed independently of any individual eBay auction, ruling out mechanical correlation between the normalisation and treatment assignment.

The baseline specification is

$$\frac{\text{SalePrice}_i}{\text{BookValue}_i} = \beta_0 + \beta_1 \mathbb{1}(\text{Shill}_i) + \beta_2 \text{GradeFE}_i + \epsilon_i, \quad (1)$$

where $\mathbb{1}(\text{Shill}_i)$ equals 1 if auction i is classified as suspicious by the clustering procedure and GradeFE_i is a full set of coin-grade fixed effects. The preferred specification adds seller fixed effects and restricts the sample to the 35 sellers (of 172) who run both flagged

and unflagged auctions; the identifying variation is therefore within seller. The remaining rows in Table 12 address the other identification threats discussed in Section 4.5: item-quality controls (grade fixed effects, $\log n_{\text{bidders}}$), reweighting on observables (propensity-score matching with and without seller absorption), and increasingly tight exact-match counterfactuals at the seller-and-grade and seller-and-coin levels.

Table 12: Effect of Shill Bidding on Sale Premium

Estimator	Coef.	SE	p-value	95% CI	N	Addresses
OLS	0.165	0.082	0.047	[0.002, 0.327]	1,558	–
OLS + Grade FE	0.174	0.071	0.016	[0.034, 0.315]	1,558	item heterog.
+ $\log(n_{\text{bidders}})$	0.180	0.069	0.010	[0.044, 0.316]	1,558	competition
+ Seller FE	0.224	0.058	0.0001	[0.110, 0.338]	1,558	seller heterog.
Within-seller subsample [†]	0.228	0.055	<0.0001	[0.120, 0.336]	1,231	within-seller var.
PSM (grade + book value)	0.166	0.082	0.043	[0.006, 0.326]	1,558	joint observable selection
PSM + Seller FE in p-score	0.252	0.063	0.0001	[0.128, 0.375]	140 prs	+ seller heterog.
NN match (exact coin) [‡]	0.120	0.063	0.055	[−0.003, 0.243]	95 prs	within-coin
Exact: seller × grade	0.184	0.051	0.0003	[0.084, 0.284]	114 prs	within-seller-and-grade
Exact: seller × grade × country × KM	0.034	0.043	0.42	[−0.050, 0.118]	65 prs	within identical coin

[†] Restricted to the 35 sellers running both flagged and unflagged auctions. [‡] Same country, KM#, grade, and book value; coins may differ in other unobserved characteristics. OLS / FE: SEs clustered by country (130 clusters). PSM: bootstrap SE. Exact match: paired SE. Bold indicates the preferred specification.

Reading the gradient

The point estimates fall into three groups, with the gradient across groups offering prima facie evidence of bounded selection bias rather than a fragile single-design result. Specifications that control only for observable item characteristics (OLS, Grade FE, $\log n_{\text{bidders}}$, PSM, exact-coin NN) give estimates between 0.12 and 0.18, all positive, with statistical significance at conventional levels except the tightest exact-coin match ($p = 0.055$). Specifications that absorb seller heterogeneity (OLS + Grade FE + Seller FE, the within-seller subsample, PSM with Seller FE in the propensity score, and the seller-and-grade exact match) land between 0.18 and 0.25, with substantially tighter standard errors and p-values below 0.001. The single specification that conditions on the exact same coin from the exact same seller produces a small and statistically insignificant estimate (0.034, $p = 0.42$) on a sample of only 65 matched pairs.

Controlling for seller heterogeneity *increases* the estimated premium rather than eliminating it. The within-seller comparison (the same seller’s shill-flagged auctions versus that same seller’s non-flagged auctions) yields the largest and most precise effect, and ipso facto rules out time-invariant seller heterogeneity as an alternative explanation. This is inconsistent with a story in which shill-flagged auctions disproportionately come from

sellers who already command higher prices for unrelated reasons; if anything, seller-level selection has been masking the effect downward.

Table 12 shows that the gap between the between-seller and within-seller estimates is due primarily to seller heterogeneity rather than the sample restriction (adding seller fixed effects on the full sample raises the estimate from 0.174 to 0.224; the restriction to the 35 mixed-portfolio sellers moves it only to 0.228). Flagged auctions disproportionately originate from sellers whose auctions otherwise fetch lower prices relative to book value, conditional on grade. This can likely be attributed to a quality difference that is unobserved in the data, but observable by buyers. Such information could be signalled in listing images or item descriptions, neither of which enters the dataset.

The tightest exact-coin-within-seller match, however, collapses to a small and noisy estimate, suggesting that the within-seller premium is concentrated in comparisons across (rather than within) coin types. Power is limited: only 65 shill-flagged auctions admit a same-seller, same-coin counterpart in the data, and we cannot reject either a small premium or a null effect at this match level. The upper end of the confidence interval (0.118) nonetheless remains economically meaningful and broadly consistent with the other specifications.

Robustness to unobservable selection: Oster bounds

To gauge how strong unobservable selection would have to be to overturn the result, I follow Oster (2019). Let $\tilde{\beta}, \tilde{R}$ be the coefficient and R^2 from the uncontrolled OLS regression and $\hat{\beta}, \hat{R}$ those from the OLS + Grade FE + Seller FE specification. Define $R_{\max}$ as a hypothetical R^2 achievable if all unobservables were observed; Oster recommends $R_{\max} = 1.3\hat{R}$ as a benchmark.

The bias-adjusted coefficient $\beta^*(\delta = 1, R_{\max})$ assumes that selection on unobservables is proportional to selection on observables. The threshold $|\delta|$ at which β^* would equal zero indicates how much stronger unobservable selection must be (relative to observable) to nullify the result.

For my data: $\tilde{\beta} = 0.164$, $\tilde{R} = 0.001$, $\hat{\beta} = 0.224$, $\hat{R} = 0.469$. At $R_{\max} = 1.3\hat{R} = 0.609$, $|\delta| = 12.5$ and $\beta^*(\delta = 1) = 0.242$. If unobservable selection is proportional in magnitude to observable selection, then for the result to vanish unobservables must dominate observables; this, however, is not what the data show. First, $|\delta| \gg 1$ across reasonable values of $R_{\max}$ (3.7 at $R_{\max} = 0.94$, 12.5 at the Oster benchmark): unobservable selection would need to be at least three to twelve times stronger in magnitude than observable selection to drive the estimate to zero, and a fortiori the result is robust at less demanding $R_{\max}$ benchmarks. Secondly, δ is *negative*: because Seller FE *raises* the estimated coefficient, ob-

servable selection biases the result downward. Under Oster’s proportionality assumption, unobservable selection would do the same; ergo, for unobservables to nullify the result they would need to bias it upward, opposite-signed to observables and several times larger in magnitude, an implausible configuration. The bias-adjusted $\beta^*(\delta = 1)$ exceeds the controlled estimate (0.242 vs. 0.224), suggesting that if anything the controlled estimate understates the true effect.

A caveat: the Oster framework is more reliable when observable controls absorb a moderate share of variance and have meaningful selection content. Here $\tilde{R}$ is very small (0.001), and $\hat{R}$ is large (0.47) primarily because Grade FE and Seller FE absorb substantial baseline variation in seller- and grade-specific premia, most of which is unrelated to skill selection per se. The Oster machinery applies mechanically but the proportional-selection assumption is on weaker ground when $\tilde{\beta} - \hat{\beta}$ is negative as here. I therefore present Oster as a supporting robustness check, not as the primary identification strategy. The within-seller comparison (Section 4.5) is the cleanest available identification and does not rely on Oster’s proportionality.

Economic magnitude

A 17–23 percentage-point premium on book value works out to \$17–23 on a \$100 coin and \$28–38 on a coin at the sample mean book value of \$165. This is the buyer’s loss per flagged auction and, conversely, the seller’s gain.

5 Conclusion

This paper has endeavoured to shew that shill bidding can be detected from observable bidding behaviour and that the resulting classification supports a substantial estimate of the price premium, identified off variation within seller.

The clustering procedure recovers a single suspicious cluster containing 9.2% of completed sales, dominated by the price-squeezing trace described by Engelberg and Williams (2009): Successive Proxy Outbidding is essentially absent from normal bidders and pervasive in the flagged cluster, and Seller Loyalty alongside the other cross-cutting markers reinforces a coordination interpretation. The 9.2% auction-level prevalence is best read as a conservative lower bound. Engelberg and Williams (2009)’s bid-level estimate of 1.39%, scaled to the auction level under a Poisson approximation on the sample’s mean of 22.7 bids per auction, implies roughly 27% of auctions containing at least one shill bid. The clustering captures only the clearest cases; more sophisticated shills who mimic legitimate

behaviour are missed.

The price-premium estimate from the preferred within-seller specification is 22.8 percentage points of book value ($p < 0.0001$, $n = 1,231$), comfortably inside the 16–44% range Nikitkov and Bay (2015) documents in field experiments. The substantive contribution of the within-seller comparison is not the point estimate alone but what the gradient across designs reveals: absorbing seller heterogeneity *raises* the estimate rather than eliminating it, indicating that observable seller selection has been biasing reduced-form estimates downward in the prior literature. Oster (2019) bounds rule out proportional-selection stories of comparable magnitude.

Three limitations bear on the interpretation. First, the absence of ground-truth validation means I cannot compute standard classification metrics; the clustering may admit false positives and miss sophisticated shills who mimic legitimate patterns. Measurement error in the treatment indicator attenuates the regression coefficient towards zero, so the within-seller estimate should be read as a lower bound on the true effect. Secondly, the sample is one product category on one platform; collectible coins are unusually well-suited to within-seller comparison and to exact-match designs, but generalisation to other eBay categories, let alone to platforms outwith eBay, requires extending the empirical strategy with category-appropriate book values. Thirdly, the tightest exact-coin match within seller (65 pairs) cannot distinguish a small premium from a null effect at conventional significance levels; withal, the broader within-seller comparison and the seller-and-grade exact match are the cleanest specifications that retain adequate power.

Because the detection metrics do not use auction-price information, they can in principle be applied in real time before an auction closes. The platform-governance implication is that the same hierarchical clustering used here is feasible against the live bid stream, an option that prior identification-based approaches (Marshall, 2024) do not afford. Calibration against confirmed shill cases would be required before operational use, but the data and the methodology are in place.

References

- Abadie, A., & Imbens, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74(1), 235–267.
- Alzahrani, A., & Sadaoui, S. (2018). Scraping and preprocessing commercial auction data for fraud classification. *arXiv preprint arXiv:1806.00656*.
- Alzahrani, A., & Sadaoui, S. (2020). Clustering and labeling auction fraud data. In *Data management, analytics and innovation* (pp. 269–283). Springer.
- Anowar, F., Sadaoui, S., & Mouhoub, M. (2018). Auction fraud classification based on clustering and sampling techniques. In *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)* (pp. 366–371). doi: 10.1109/ICMLA.2018.00061
- Bajari, P., & Hortacısu, A. (2003). The winner’s curse, reserve prices, and endogenous entry: Empirical insights from eBay auctions. *RAND Journal of Economics*, 34(2), 329–355.
- Barbaro, S., & Bracht, B. (2021). Shilling, squeezing, sniping. a further explanation for late bidding in online second-price auctions. *Journal of Behavioral and Experimental Finance*, 31, 100553.
- Chakraborty, I., & Kosmopoulou, G. (2004). Auctions with shill bidding. *Economic Theory*, 24(2), 271–287.
- eBay Inc. (2024, February). *eBay Inc. reports fourth quarter and full year 2023 results*. Press release, PR Newswire.
- Engelberg, J., & Williams, J. (2009). eBay’s proxy bidding: A license to shill. *Journal of Economic Behavior & Organization*, 72(1), 509–526.
- Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8), 861–874.
- Guerre, E., Perrigne, I., & Vuong, Q. (2000). Optimal nonparametric estimation of first-price auctions. *Econometrica*, 68(3), 525–574.
- Guha, S., Rastogi, R., & Shim, K. (1998). CURE: an efficient clustering algorithm for large databases. In *Proceedings of the 1998 ACM SIGMOD international conference on management of data* (pp. 73–84).
- Hickman, B. R. (2010). On the pricing rule in electronic auctions. *International Journal of Industrial Organization*, 28(5), 423–433.
- Hickman, B. R., Boehnke, J., & Bodoh-Creed, A. L. (2021). How efficient are decentralized auction platforms? *The Review of Economic Studies*, 88(1), 91–125.
- Hickman, B. R., Hubbard, T. P., & Paarsch, H. J. (2017). Identification and estimation of a bidding model for electronic auctions. *Quantitative Economics*, 8(1), 223–270.

- Ingebretsen Carlson, J., & Wu, T. (2022). Shill bidding and information in eBay auctions: A laboratory study. *Journal of Economic Behavior & Organization*, 202, 341–360.
- Kauffman, R. J., & Wood, C. A. (2003). Running up the bid: detecting, predicting, and preventing reserve price shilling in online auctions. In *Proceedings of the 5th international conference on electronic commerce* (pp. 259–265).
- Kosmopoulou, G., & De Silva, D. G. (2007). The effect of shill bidding upon prices: Experimental evidence. *International Journal of Industrial Organization*, 25(2), 291–313.
- Lamy, L. (2009). The shill bidding effect versus the linkage principle. *Journal of Economic Theory*, 144(1), 390–413.
- MacKinnon, J. G., Nielsen, M. Ø., & Webb, M. D. (2023). Cluster-robust inference: A guide to empirical practice. *Journal of Econometrics*, 232(2), 272–299.
- Marshall, G. (2024). Identification in English auctions with shill bidding. *Quantitative Marketing and Economics*, 22, 193–222.
- McCannon, B. C., & Minuci, E. (2020). Shill bidding and trust. *Journal of Behavioral and Experimental Finance*, 26, 100279.
- Milgrom, P. (2004). *Putting auction theory to work*. Cambridge: Cambridge University Press.
- Nikitkov, A., & Bay, D. (2015). Shill bidding: Empirical evidence of its effectiveness and likelihood of detection in online auction systems. *International Journal of Accounting Information Systems*, 16, 42–54.
- Ockenfels, A., & Roth, A. E. (2006). Late and multiple bidding in second price internet auctions: Theory and evidence concerning different rules for ending an auction. *American Economic Review*, 96(4), 1093–1103.
- Oster, E. (2019). Unobservable selection and coefficient stability: Theory and evidence. *Journal of Business & Economic Statistics*, 37(2), 187–204.
- Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1), 41–55.
- Roth, A. E., & Ockenfels, A. (2002). Last-minute bidding and the rules for ending second-price auctions: Evidence from eBay and Amazon auctions on the internet. *American Economic Review*, 92(4), 1093–1103.
- Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20, 53–65.
- Vickrey, W. (1961). Counterspeculation, auctions, and competitive sealed tenders. *The Journal of Finance*, 16(1), 8–37.